System rekomendacji książek oparty na grafie użytkownik–książka

Antczak Jakub [1] Baczyńska Justyna [1] Gromski Wojciech [1] Łubniewska Maria [1]

[1]Politechnika Wrocławska

Motywacja

W dobie ogromnych baz danych z książkami użytkownicy często mają trudność z wyborem interesujących tytułów. Systemy rekomendacyjne są kluczowym elementem platform z książkami, pomagając w odkrywaniu nowych pozycji dopasowanych do preferencji odbiorcy.

W ramach projektu wykorzystaliśmy zbiór 228 milionów recenzji książek do budowy systemu rekomendacji opartego na analizie grafu dwudzielnego użytkownik–książka. Relacje między węzłami (czyli recenzje) pozwoliły na wyznaczenie podobieństw między książkami, co umożliwiło proponowanie użytkownikom nowych pozycji na podstawie wspólnego sąsiedztwa i prostych miar grafowych.

Celem było zbudowanie rozwiązania opartego na klasycznych metodach rekomendacji w sieci złożonej.

Wstępna analiza danych

Dane interakcji między użytkownikami a książkami pochodzą z serwisu Goodreads i początkowo obejmowały 228 648 342 rekordy. Każdy wiersz zawierał informacje o identyfikatorze użytkownika, książki, statusie przeczytania, ocenie oraz obecności recenzji.

W pierwszym kroku usunięto interakcje, w których książka nie została oznaczona jako przeczytana. Następnie, aby skupić się na pozytywnych rekomendacjach, zachowano jedynie te przypadki, w których użytkownik wystawił ocenę co najmniej 4 gwiazdki. W kolejnym etapie odfiltrowano użytkowników, którzy przeczytali mniej niż 20 książek oraz książki ocenione przez mniej niż 20 użytkowników.

Po przetworzeniu danych finalny zbiór objął 93 622 użytkowników i 106 595 książek, co pozwoliło na zbudowanie bardziej reprezentatywnego grafu interakcji.

Budowa grafu

Modelujemy relacje czytelnik–książka jako **graf dwudzielny** (ang. *bipartite graph*) $G = (U \cup B, E)$, gdzie U to użytkownicy, B to książki, a krawędź $(u, b) \in E$ oznacza, że użytkownik u przeczytał książkę b.

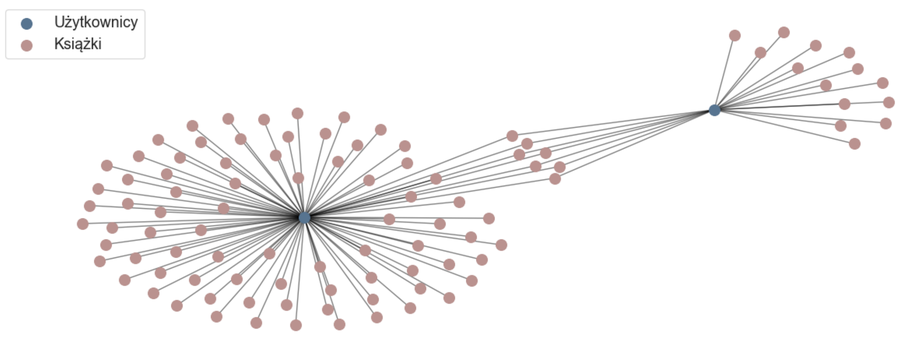
Dla dowolnej książki $b \in B$ definiujemy zbiór użytkowników

$$U(b) = \{\, u \in U : (u, b) \in E \},$$

a dla dowolnego użytkownika $u \in U$ zbiór przeczytanych książek

$$B(u) = \{\, b \in B : (u, b) \in E \}.$$

Po utworzeniu graf posiadał 200 217 wierzchołków i 8 272 041 wierszy.

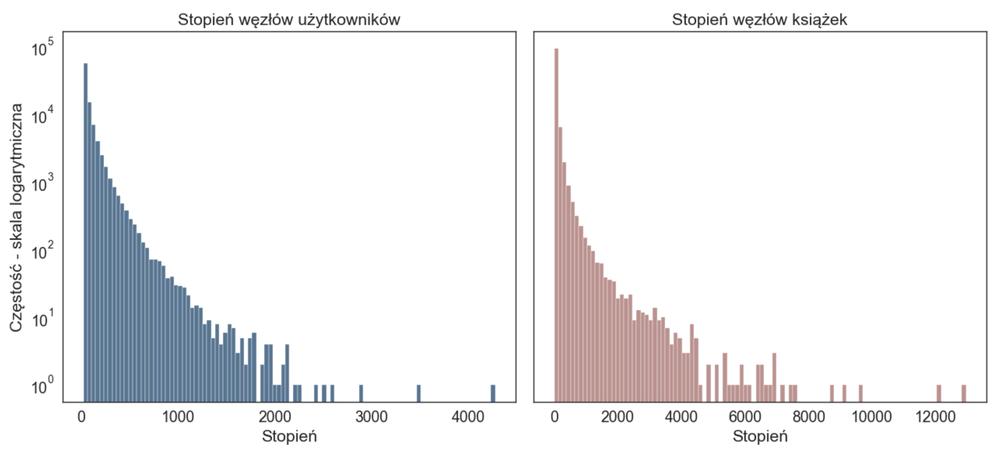


Rysunek 1. Fragment grafu przedstawiający dwóch użytkowników oraz ocenione przez nich książki.

Analiza grafu

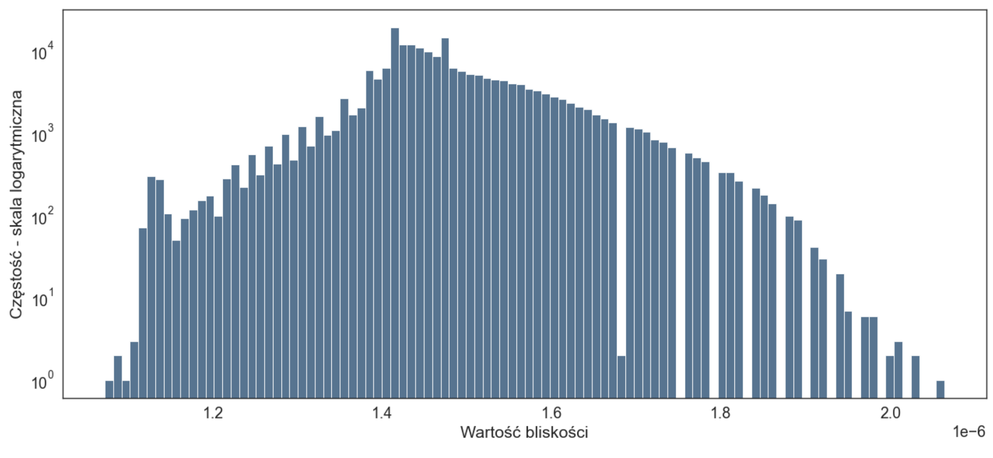
Średnica grafu wynosi 4. Oznacza to, że najdłuższa z najkrótszych ścieżek w grafie ma dokładnie 4 krawędzie. W dwudzielnym grafie, jeżeli średnica jest liczbą parzystą musi się kończyć w tych samych częściach, a więc najdalsza para składa się z tego samego rodzaju wierzchołków.

Z obu wykresów stopni węzłów na Rysunku 2 możemy zauważyć wyraźną prawoskośność obu zbiorów. Zdecydowana większość czytelników sięga po niewielką liczbę książek, ale istnieją również nieliczni ponad przeciętnie aktywni. Analogicznie - większość książek ma mały odsetek czytelników, a bestsellery gromadzą dużą liczbę użytkowników.



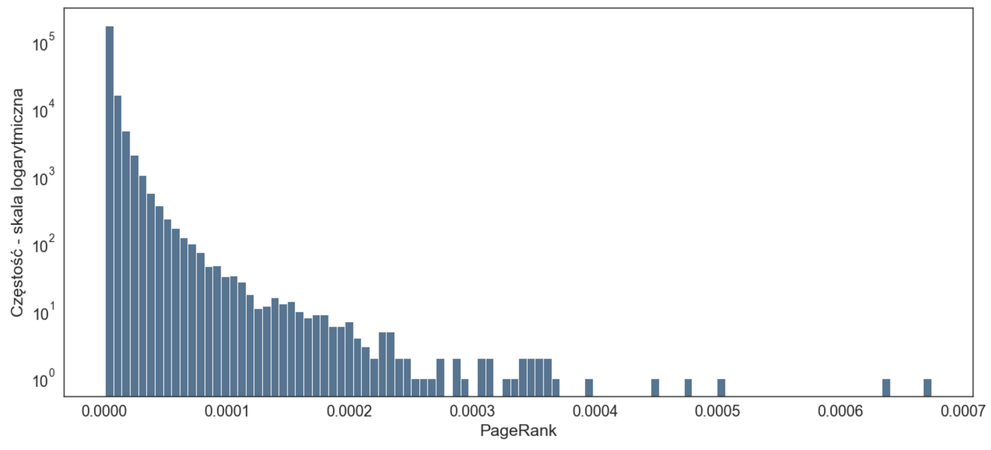
Rysunek 2. Histogramy stopni węzłów użytkowników oraz książek.

Z wykresu na Rysunku 3 widać, że większość węzłów koncentruje się wokół środkowych wartości bliskości, co świadczy o sieci z równomiernym dostępem do pozostałych węzłów. Wierzchołki o najwyższych wartościach pełnią funkcję *hubów* – to najbardziej aktywni czytelnicy oraz bestsellery. Po drugiej stronie znajdują się węzły o najniższej bliskości, reprezentują one niszowe tytuły lub mało aktywnych użytkowników.



Rysunek 3. Histogram wartości bliskości (ang. *closeness centrality*).

Histogram rozkładu wartości PageRank, widoczny na Rysunku 4, w sieci pokazuje silną prawoskośność. Duża liczba węzłów ma bliskie zeru wartości, jedynie kilka z nich działa jako kluczowe *huby*. Oznacza to, że w sieci dominują mało popularne tytuły i mało aktywni czytelnicy.



Rysunek 4. Histogram PageRank.

Metody rekomendacji

Filtracja kolaboracyjna (ang. *Collaborative filtering*)

Aby ocenić podobieństwo dwóch książek $b_i, b_j \in B$, wykorzystujemy cztery miary oparte na wspólnych użytkownikach.

- Indeks Jaccarda

$$s_{\mathrm{J}}(b_i, b_j) = \frac{|U(b_i) \cap U(b_j)|}{|U(b_i) \cup U(b_j)|}$$

- Współczynnik nakładania (ang. *Overlap coefficient*)

$$s_{\mathrm{O}}(b_i, b_j) = \frac{|U(b_i) \cap U(b_j)|}{\min\big(|U(b_i)|, |U(b_j)|\big)}$$

- Indeks Adamic–Adar [1]

$$s_{\mathrm{AA}}(b_i, b_j) = \sum_{u \in U(b_i) \cap U(b_j)} \frac{1}{\ln\big(|B(u)|\big)}$$

- Alokacja zasobów (ang. *Resource Allocation*, [2])

$$s_{\mathrm{RA}}(b_i, b_j) = \sum_{u \in U(b_i) \cap U(b_j)} \frac{1}{|B(u)|}$$

Algorytm rekomendacji dla użytkownika u^*

1. Wybieramy zbiór przeczytanych przez u^* książek: $R(u^*)$.
2. Wybieramy zbiór kandydatów:
$$C = \bigcup_{b \in R(u^*)} \big(U(b) \setminus \{u^*\}\big).$$
3. Dla każdego $c \in C$ obliczamy $\mathrm{score}(u^*, c) = \sum_{b \in R(u^*)} s(b, c)$.
4. Wybieramy top-N rekomendacji na podstawie wartości score.

Spersonalizowany PageRank (ang. *Personalized PageRank*, [3])

Dla docelowego użytkownika u^* definiujemy wektor personalizacji $\mathbf{p} = \mathbf{e}_{u^*}$.

Spersonalizowany PageRank $\mathbf{r}$ jest rozwiązaniem równania:

$$\mathbf{r} \;=\; \alpha\, \mathbf{P}^T \mathbf{r} \;+\; (1 - \alpha)\, \mathbf{p},$$

gdzie α to *damping factor* (zwykle 0.85), a $\mathbf{P}$ to macierz przejścia grafu G.

Algorytm rekomendacji dla użytkownika u^*:

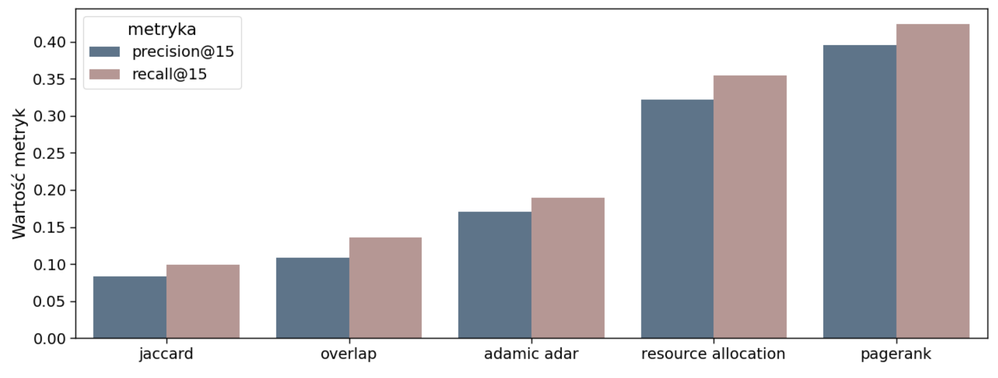
1. Obliczamy spersonalizowany PageRank $\mathbf{r}$ na grafie $G = (U \cup B, E)$, z parametrami α i $\mathbf{p}$.
2. Wybieramy zbiór przeczytanych przez u^* książek:
$$R(u^*) = \{\, b \in B : (u^*, b) \in E \}.$$
3. Wybieramy zbiór kandydatów:
$$C = B \setminus R(u^*).$$
4. Dla każdego kandydata $b \in C$ definiujemy
$$\mathrm{score}(u^*, b) = r(b),$$
czyli wartość PageRank węzła odpowiadającego książce b.
5. Wybieramy top-N książek o największych wartościach score.

Porównanie metod rekomendacji

Wybrano 150 użytkowników. Dla każdego użytkownika usunięto losowo 20% książek, z którymi wcześniej wchodził w interakcję. Następnie, dla każdego z nich wygenerowano 15 rekomendacji. Na tej podstawie obliczono wybrane metryki skuteczności rekomendacji.

Zastosowane metryki:

- **precision@k** – odsetek rekomendowanych książek (z top-k), które faktycznie należały do usuniętych (czyli były trafne),
- **recall@k** – odsetek usuniętych książek, które znalazły się wśród rekomendowanych (czyli zostały odzyskane).



Rysunek 5. Porównanie metryk dla różnych metod rekomendacji.

Na wykresie na Rysunku 5 przedstawione są uśrednione wartości metryk precision@15 i recall@15 dla każdej z testowanych metod. Najlepsze wyniki osiągnęła metoda PageRank, charakteryzując się najwyższą wartością zarówno precision, jak i recall. Najsłabiej wypadł algorytm Jaccard, uzyskując najniższe wartości w obu metrykach.

Bibliografia

[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[2] H. Li, P. Liang, and J. Hu. A network resource allocation recommendation method with an improved similarity measure. *arXiv preprint arXiv:2307.03399*, 2023.

[3] C. Musto, P. Lops, M. de Gemmis, and G. Semeraro. Context-aware graph-based recommendations exploiting personalized pagerank. *Knowledge-Based Systems*, 216:106806, 2021.