Data mining як засіб структурування великих обсягів даних (Big Data) та їх обробки

Зміст

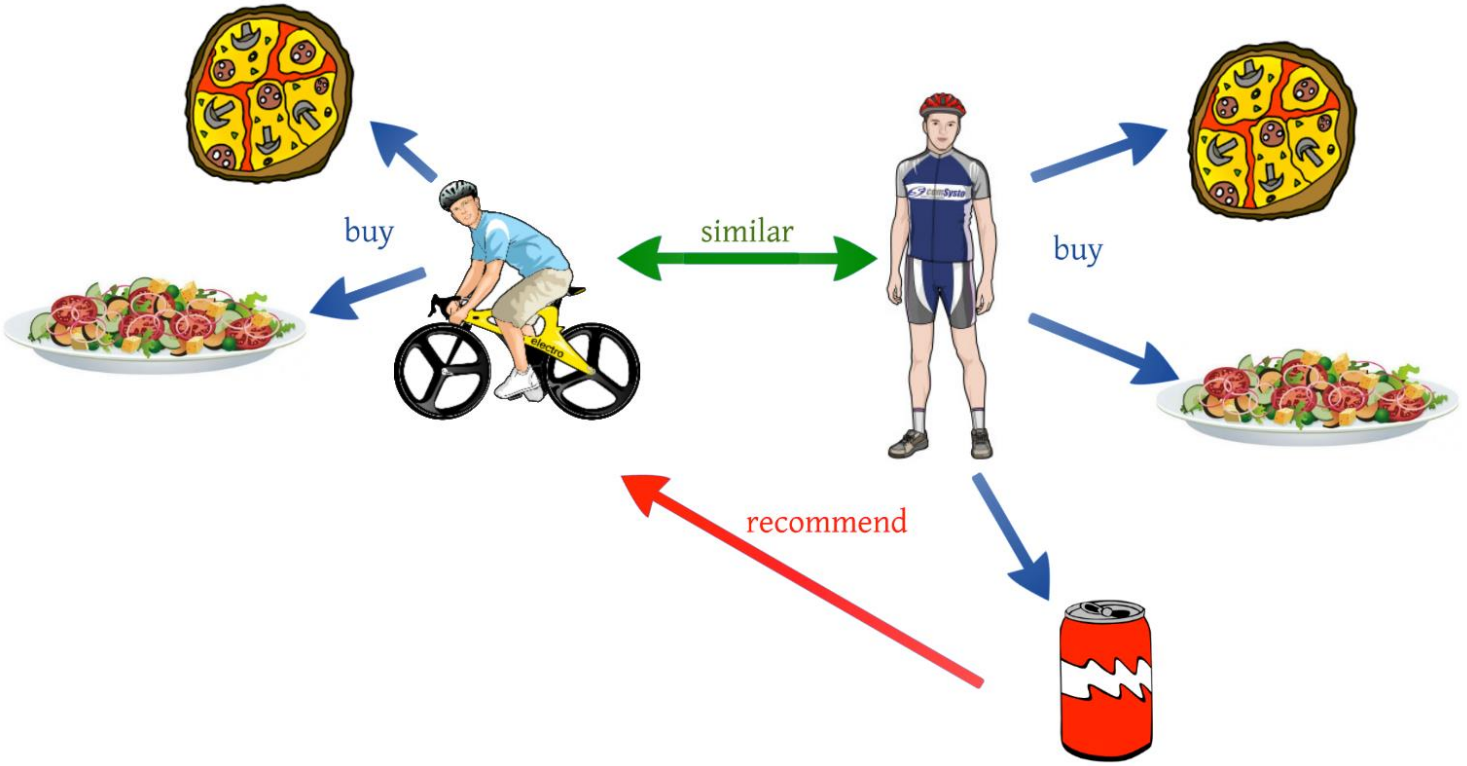
- Big Data & Рекомендаційні системи
- Соціальні зв’язки – джерело даних для РС
- Постановка гіпотези
- Експерементальні дані
- Результати

Big Data & Рекомендаційні системи

- **Рекомендаційна система** (РС) — підклас системи фільтрації інформації, яка будує рейтинговий перелік об'єктів (фільми, музика, книги, новини, веб-сайти, події), яким користувач може надати перевагу.
- Типи РС:
  - Фільтрація вмісту
  - Колаборативна фільтрація

Фільтрація на основі вмісту

- Цей метод фільтрації ґрунтується на описі речей, що рекомендуються, і вподобань користувача за допомогою ключових слів.

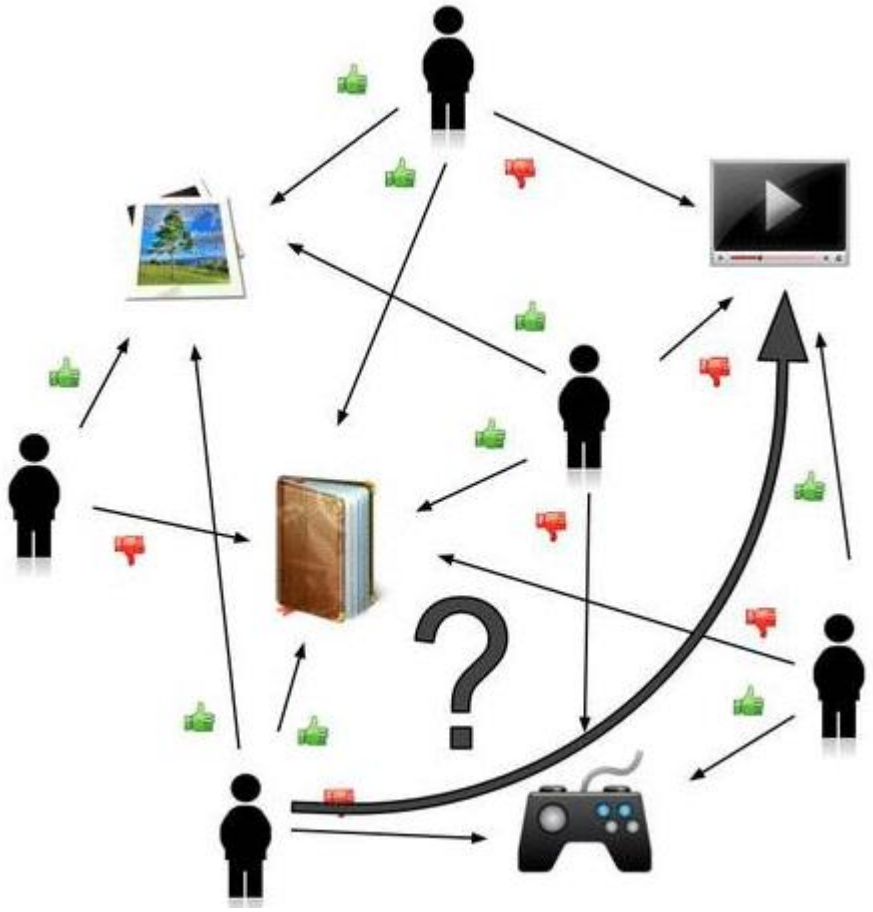


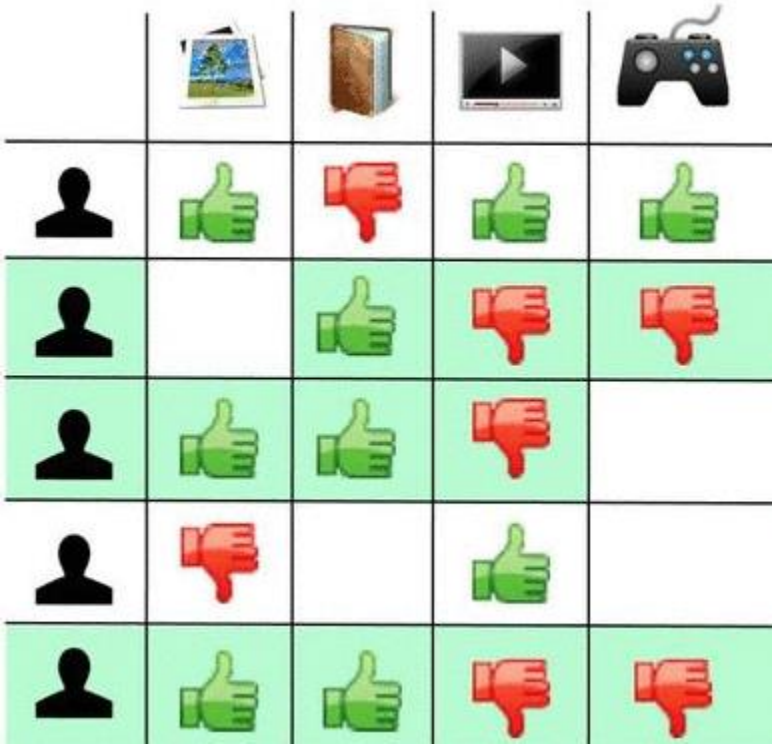
Колаборативна фільтрація

– Це один з методів побудови прогнозу, який використовує відомі оцінки групи користувачів для прогнозування невідомих оцінок іншого користувача.

$$\mathrm{simil}(x,y) = \cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{||\vec{x}|| \times ||\vec{y}||} = \frac{\sum_{i\in I_{xy}} r_{x,i} r_{y,i}}{\sqrt{\sum_{i\in I_x} r_{x,i}^2}\sqrt{\sum_{i\in I_y} r_{y,i}^2}}$$

Приклади Колаборативної фільтрації



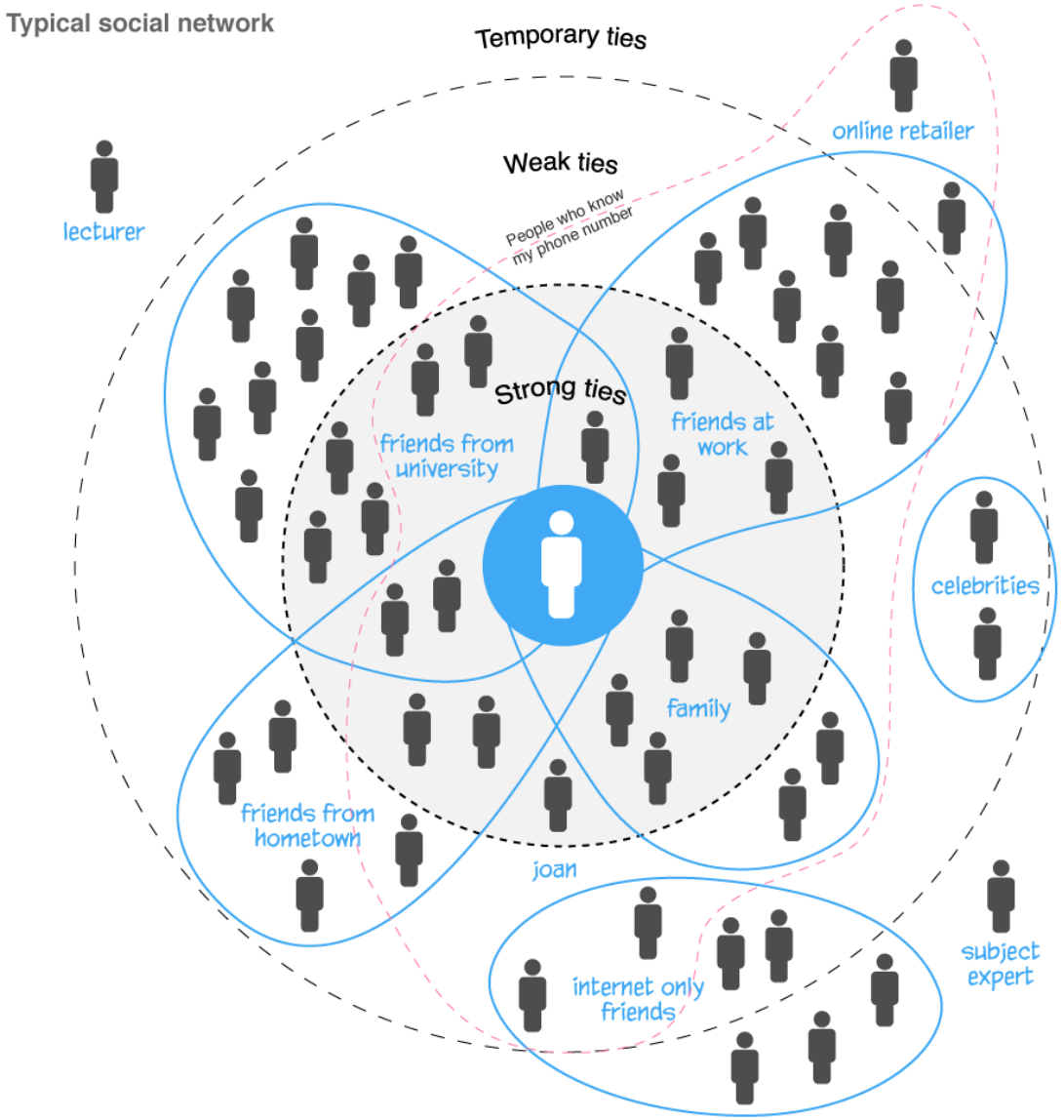


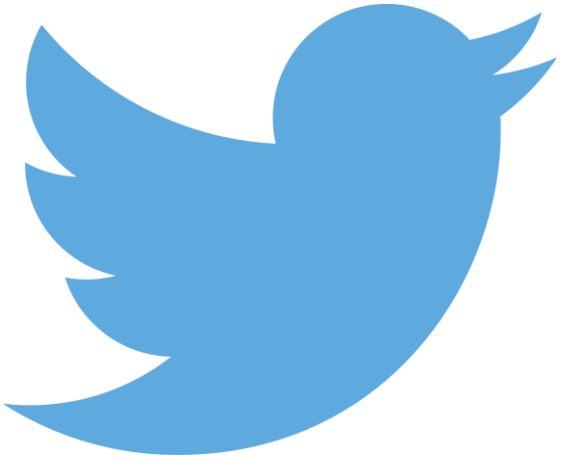
Netflix prize

- 100,480,507 оцінок фільмів
- 480,189 користувачів
- 17,770 фільмів
- 15% покращення
- 1 billion USD

Соціальні зв’язки – джерело даних для РС

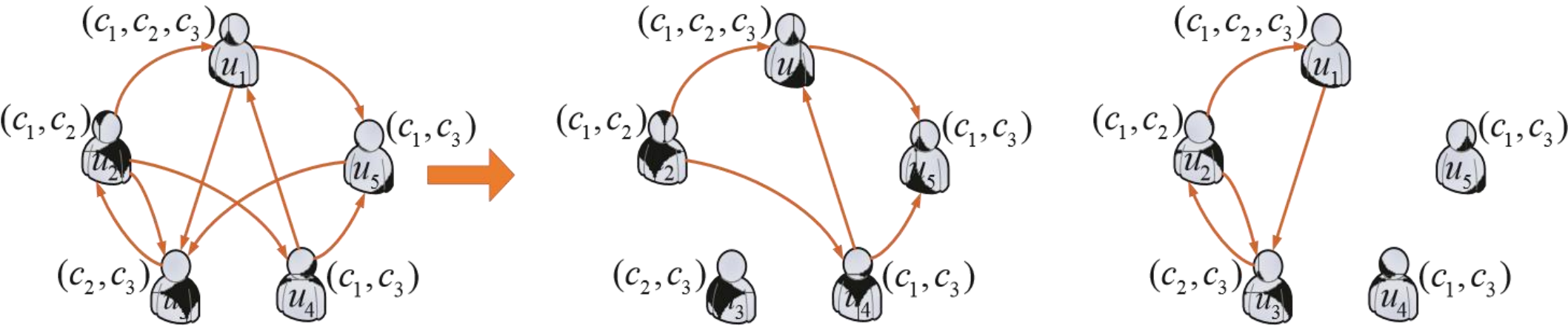
Друзі в соціальних мережах





Алгоритми аналізу друзів

- Категоризація друзів
- Оцінка схожості



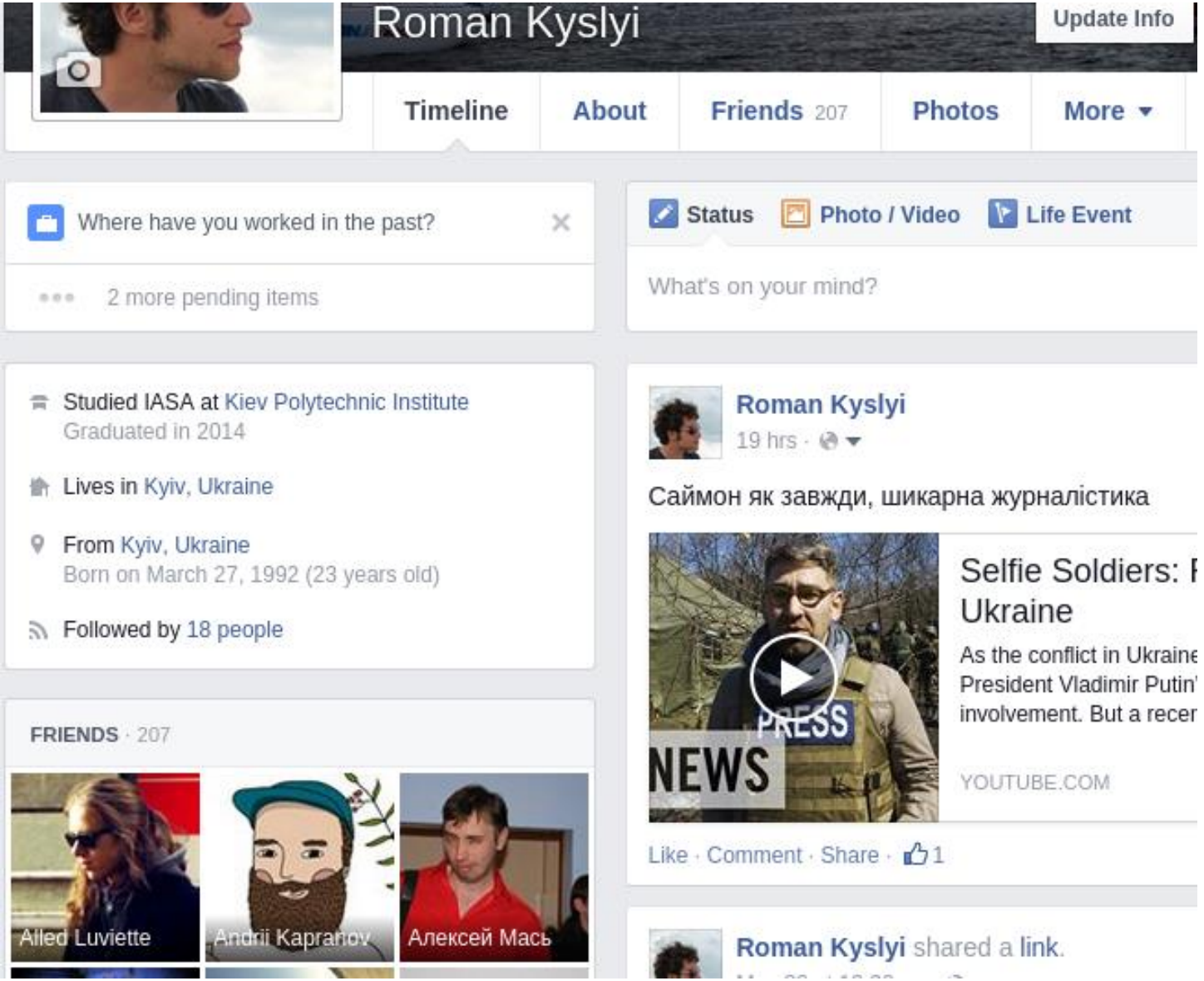
Проблема “холодного старту”

Рішення:

- Випадкова рекомендація
- Найбільш популярна рекомендація
- Прогнозована на базі соціальних мереж

Постановка гіпотези

Профіль – додаткове джерело даних



Likes

+ Add Likes

All Likes 176 | Movies | TV Shows | Music | Books | Sports Teams | Athletes | People | Restaurants | Apps and Games

SPIEGEL ONLINE
News/Media Website
✓ Liked ▾ ✓ Following

Süddeutsche Zeitung
Newspaper
✓ Liked ▾ ✓ Following

Tourclub "Globus"
Sports Team
✓ Liked ▾ ✓ Following

Серіал "Гвардія"
TV Show
✓ Liked ▾ ✓ Following

Upwork Ukraine
Company
✓ Liked ▾ ✓ Following

site.ua
Media/News/Publishing
✓ Liked ▾ ✓ Following

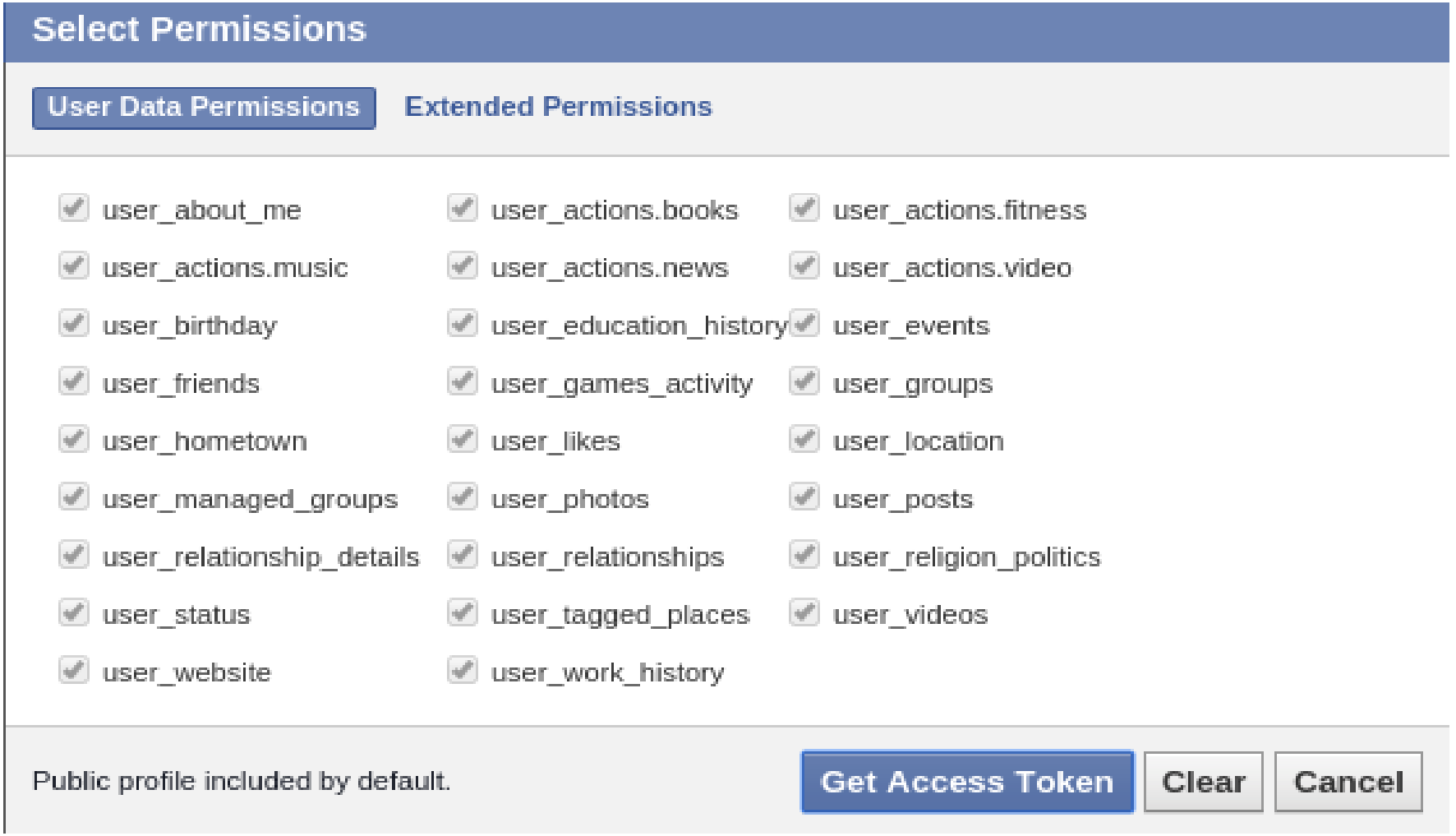
Профіль користувача


```
{
    first_name: "Roman",
    last_name: "Kyslyi",
    relationship_status: "Single",
    name: "Roman Kyslyi",
    locale: "en_US",
    gender: "male",
    verified: true,
    email: "kvrware@gmail.com",
    religion: "belive in Jedi Forse",
    birthday: "03/27/1992",
    link: "http://www.facebook.com/100001128635849",
  - location: {
        id: "111227078906045",
        name: "Kyiv, Ukraine"
    },
  + favorite_athletes: […],
  + hometown: {…},
    timezone: 3,
  - education: [
      - {
          - school: {
                id: "109195292432741",
                name: "Polytechnic High School"
            },
            type: "High School",
          - year: {
                id: "136328419721520",
                name: "2009"
            }
        },
      + {…}
    ],
    updated_time: "2014-09-24T20:02:14+0000",
    id: "100001128635849"
}
```



```
{
    first_name: "Roman",
    last_name: "Kyslyi",
    relationship_status: "Single",
    name: "Roman Kyslyi",
    locale: "en_US",
    gender: "male",
    verified: true,
    email: "kvrware@gmail.com",
    religion: "belive in Jedi Forse",
    birthday: "03/27/1992",
    link: "http://www.facebook.com/100001128635849",
  - location: {
        id: "111227078906045",
        name: "Kyiv, Ukraine"
    },
  + favorite_athletes: […],
  + hometown: {…},
    timezone: 3,
  - education: [
      - {
          - school: {
                id: "109195292432741",
                name: "Polytechnic High School"
            },
            type: "High School",
          - year: {
                id: "136328419721520",
                name: "2009"
            }
        },
      + {…}
    ],
    updated_time: "2014-09-24T20:02:14+0000",
    id: "100001128635849"
}
```

Можливості Facebook API



Модель SocialMF

- Ідея - профіль користувача має бути в середньому схожим на профіль його друзів, в межах похибки

$$\frac{1}{2}\sum_{(u,i)\text{obs.}}\left(R_{u,i}-\hat{R}_{u,i}\right)^2$$

Ваговий коефіцієнт

$$W = \sum_{N}^{i=1} (a_i * b_i) = \sum_{N}^{i=1} (a_i * d_i * f_i)$$

- ai - передбачена оцінка речі(події) і в категорії с
- di – ваговий коефіцієнт
- fi – індикатор, що показує чи присутня категорія в інтересах користувача

$$W \frac{1}{2} \sum_{(u,i)obs.} \left(R_{ui} R'_{ui} \right)^2$$

Експеременталькі дані

Набір даних Kaggle Event Recommendation Engine

- event
- user
- interested
- not_interested
- birthyear
- location
- timezone
- likes
- friends

- 90 000 подій
- 140 000 користувачів

Набір даних Epinions

- user_id
- location
- gender
- birthyear
- likes
- ratings

- 74 000 користувачів
- 590 000 оцінок
- 140 000 різних предметів

Метрика оцінювання

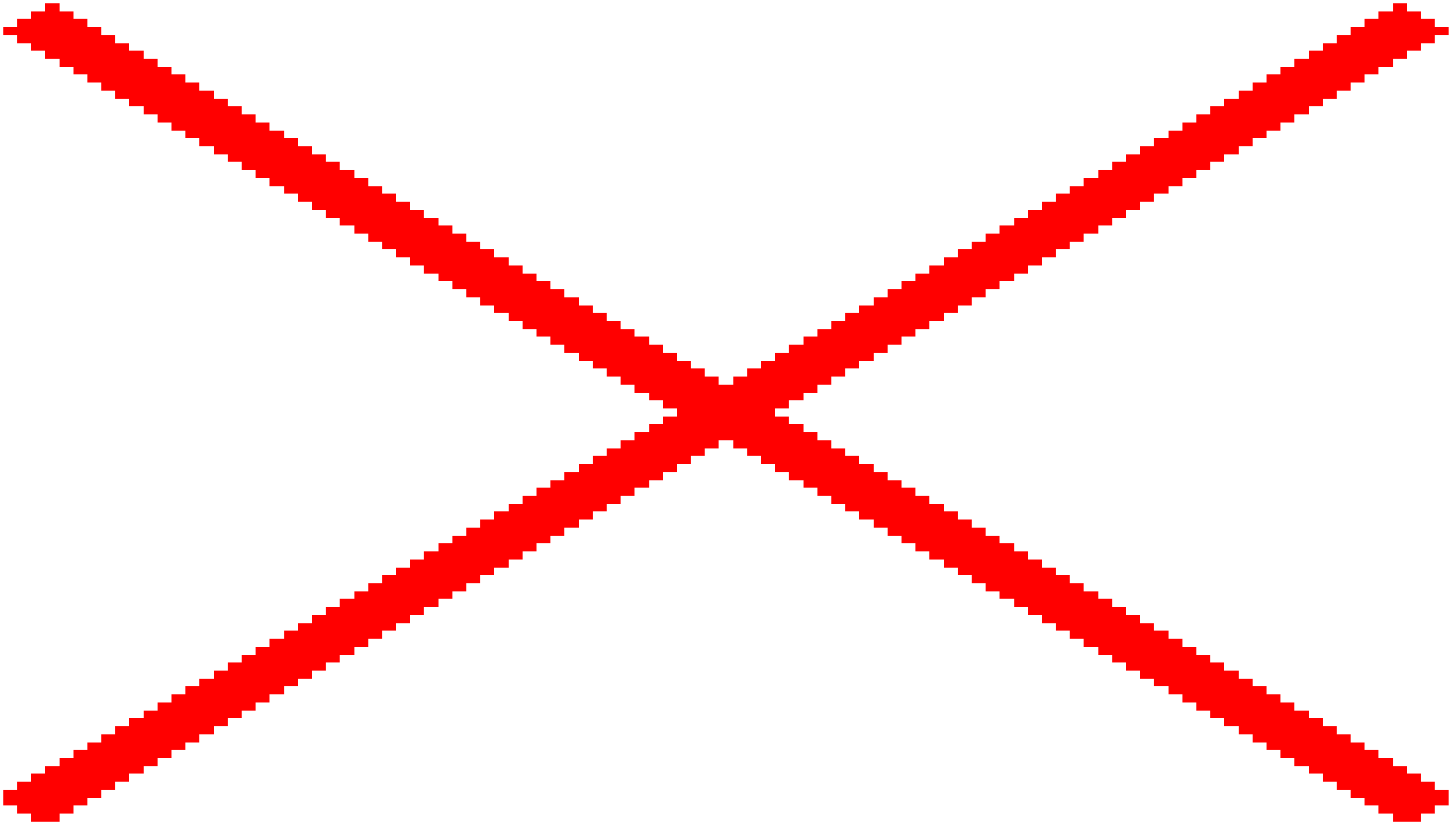
- 10 – fold cross validation

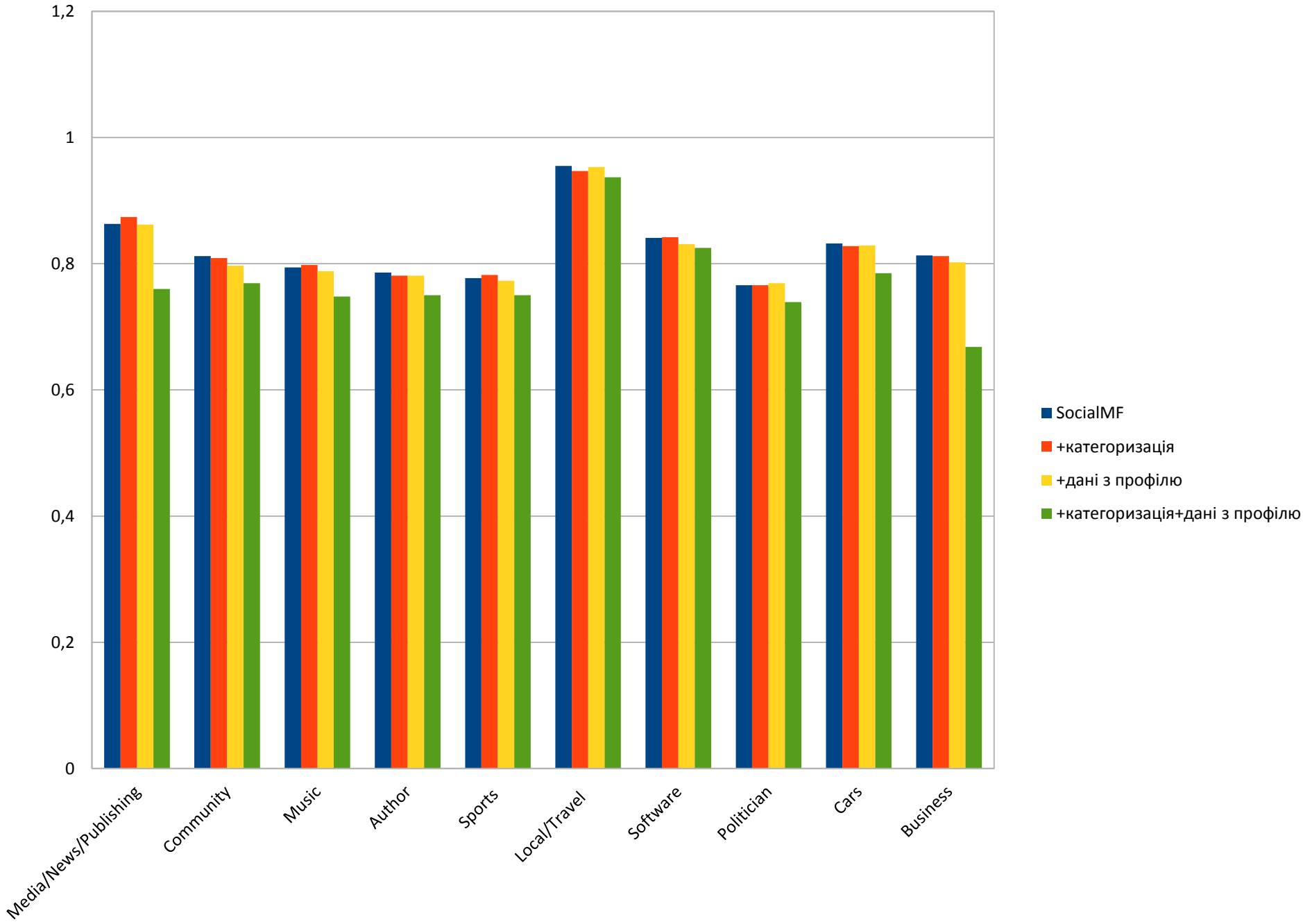
- Для оцінки точності використовувалась середньоквадратична похибка

$$RMSE = \sqrt{\frac{\sum_{(u,i)\in\mathcal{R}_{test}} (R_{u,i} - \hat{R}_{u,i})^2}{|\mathcal{R}_{test}|}}$$

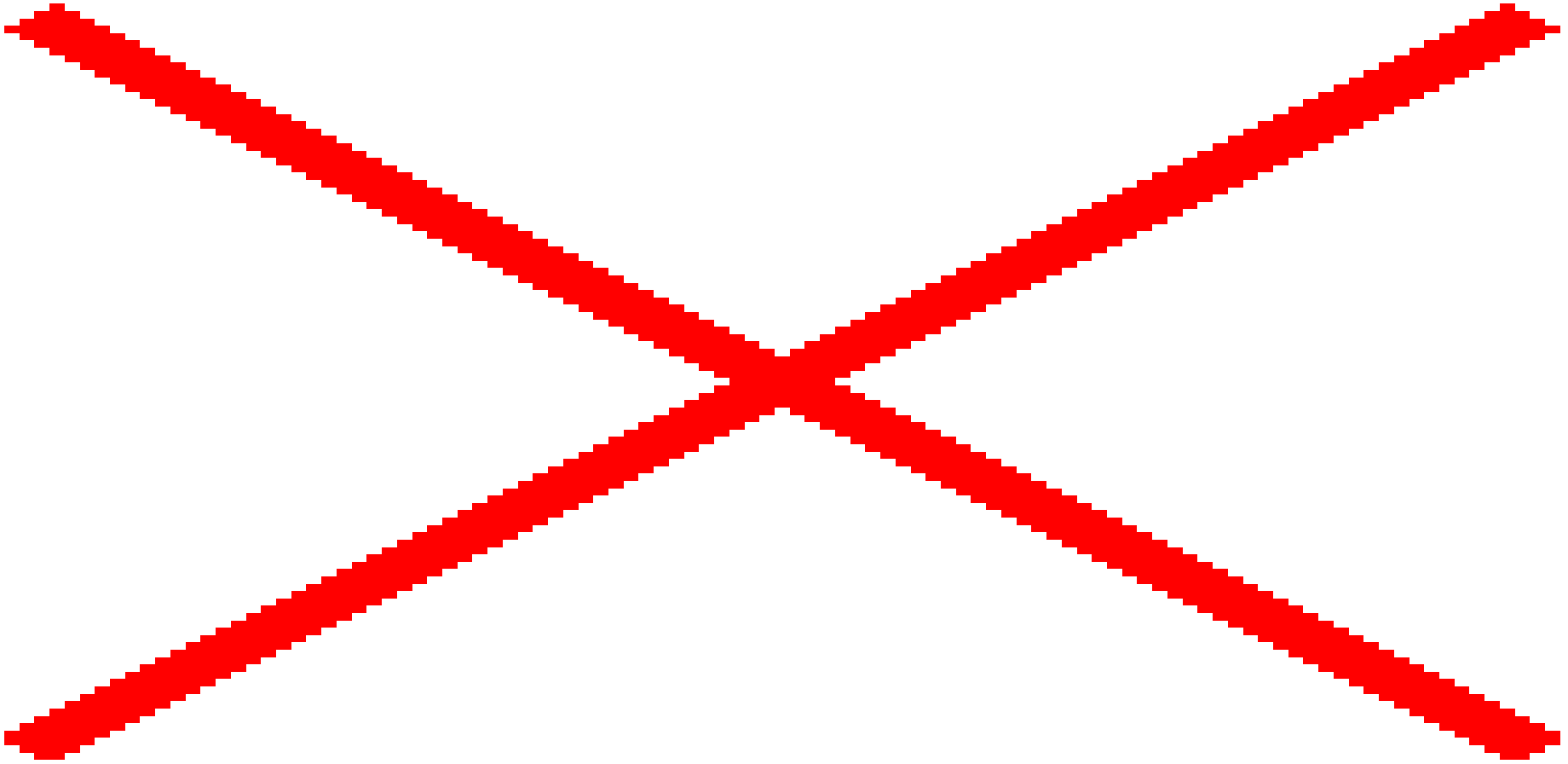
Резултати

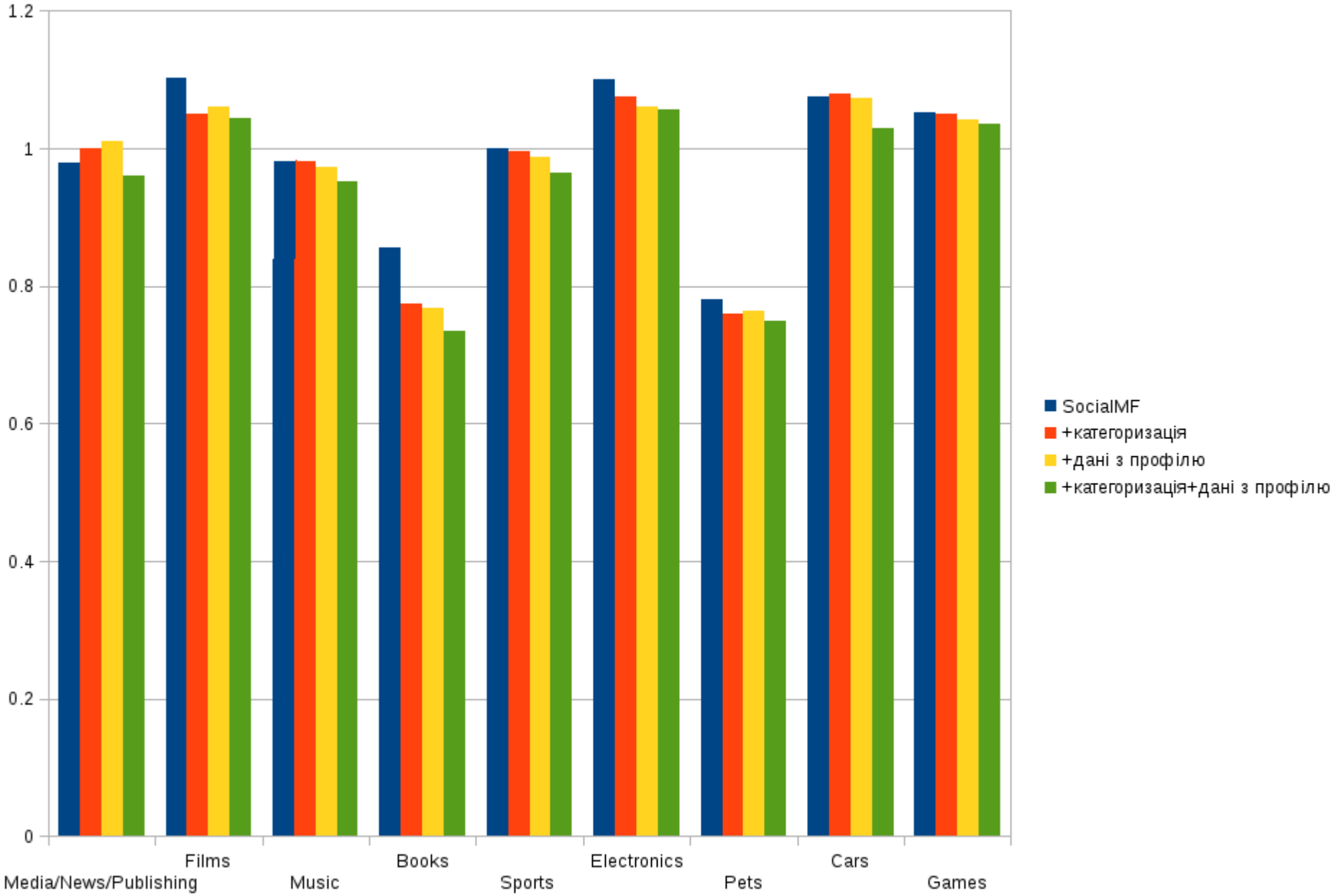
Результати тестування на наборі даних з Kaggle




1,2
1
0,8
0,6
0,4
0,2
0
Media/News/Publishing
Community
Music
Author
Sports
Local/Travel
Software
Politician
Cars
Business
SocialMF
+категоризація
+дані з профілю
+категоризація+дані з профілю

Результати тестування на наборі даних з Epinions




1.2
1
0.8
0.6
0.4
0.2
0
Media/News/Publishing
Films
Music
Books
Sports
Electronics
Pets
Cars
Games
SocialMF
+категоризація
+дані з профілю
+категоризація+дані з профілю

Питання?