ISSN 2224-087X. Електроніка та інформаційні технології. 2018. Випуск 10. С. 64–70
Electronics and information technologies. 2018. Issue 10. P. 64–70

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ В НАУКОВИХ ДОСЛІДЖЕННЯХ

УДК 004.8, 004.67

РЕКОМЕНДАЦІЙНІ ТЕХНІКИ ДЛЯ АНАЛІЗУ КРИМІНАЛІСТИЧНИХ ДАНИХ

Ю. Ейнес

Львівський національний університет імені Івана Франка,
вул. Університетська, 1, 79000 Львів, Україна
eines.yuliia@gmail.com

Видобування даних для аналізу цифрових криміналістичних даних – це окрема галузь комп'ютерних наук. Основним її завданням є вилучення шаблонів із великих даних, які використовують для аналізу під час розслідування злочинів. Одним із перспективних застосувань алгоритмів видобування даних є побудова рекомендаційних систем. Такі системи мали б пропонувати можливі майбутні напрямки розслідування.

У даній статті розглядається архітектура, метою якої є застосування технік рекомендаційних систем у криміналістиці.

Ключові слова: криміналістичний аналіз, рекомендація, колаборативна фільтрація .

Системи видобування даних для цифрового криміналістичного аналізу допомагають у поліцейських розслідуваннях, якщо вони включають масштабні колекції даних. Ця галузь комп’ютерних наук зосереджена на вилученні шаблонів із великих даних, які використовують у розслідуванні злочинів. Застосування алгоритмів такого типу має на меті спрощення розслідування для експертів в криміналістиці. Прикладом такого застосування можуть слугувати рекомендаційні системи для аналізу криміналістичних даних. Ціль цієї статті – розглянути підхід, у якому буде використано техніки рекомендаційних систем для цифрового криміналістичного аналізу. У цій статті буде обговорено використання цих технік і на великих колекціях мультимедійних даних (текст, аудіо, зображення, відео) для збільшення ефективності аналізу.

Рекомендаційні системи широко використовуються переважно у застосунках із медіа контентом (рекомендація контенту користувачам). Так як у криміналістиці використовують схожі дані, то можна помітити деякі аналогії, які можна використати. Рекомендаційні системи класифікують в залежності від особливостей, на які вони покладаються. Тобто рекомендаційні системи бувають таких типів: орієнтовані на контент (*content-based*), створені на основі колаборативної фільтрації (*collaborative filtering*) і гібридні [1]. Перший тип покладається на атрибути (технічні, семантичні, …) наявного контенту, другий – на попередні оцінки користувачів, останній – поєднання двох попередніх типів. У даній роботі буде розглянуто рекомендаційну систему, що базується на знаннях із попередніх справ для кореляції зачіпок і контексту у різних розслідуваннях.

Кореляція підказок і контекстної інформації із місця злочину є концептуальною базою для виводу правильних знань про конкретні справи. Таке знаряддя повинне

допомогти офіцерам поліції та поліцейським аналітикам виявити найімовірніші напрямки розслідування.

Архітектура системи. Запропонована архітектура складається із двох основних функціоналів, які можуть використовуватись разом або окремо. Як видно на рис. 1 ці два функціонали представляють рекомендацію зачіпок (запропонувати слідчому найкращий наступний крок для розв'язання цієї справи) і рекомендація підозрюваних [2]. Обидва функціонали базуються на алгоритмах рекомендаційних систем і на інформації, яку ми отримали із попередніх справ. Якщо розглянути це більш детально, то модель побудована на основі методів колаборативної фільтрації для «Алгоритму рекомендації зачіпок» і технік колаборативної фільтрації на базі пам'яті для «Алгоритму рекомендації підозрюваних».

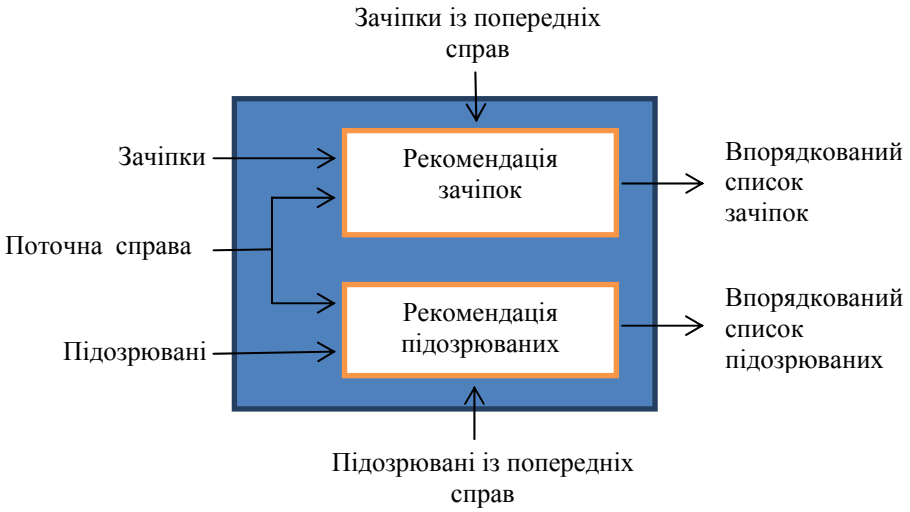

Рис.1. Функціональність системи

Алгоритм рекомендації зачіпок. Цей функціонал особливо важливий на перших етапах розслідування, коли є багато інформації і система може заощадити деякий час, автоматично створюючи перші припущення. Весь хід алгоритму зображено на рис.2. Функціонал рекомендації зачіпок починається із інформації про особливості, які характеризують справу. Ці властивості моделюються як зачіпки і слідчий може їм слідувати для покращення розслідування.

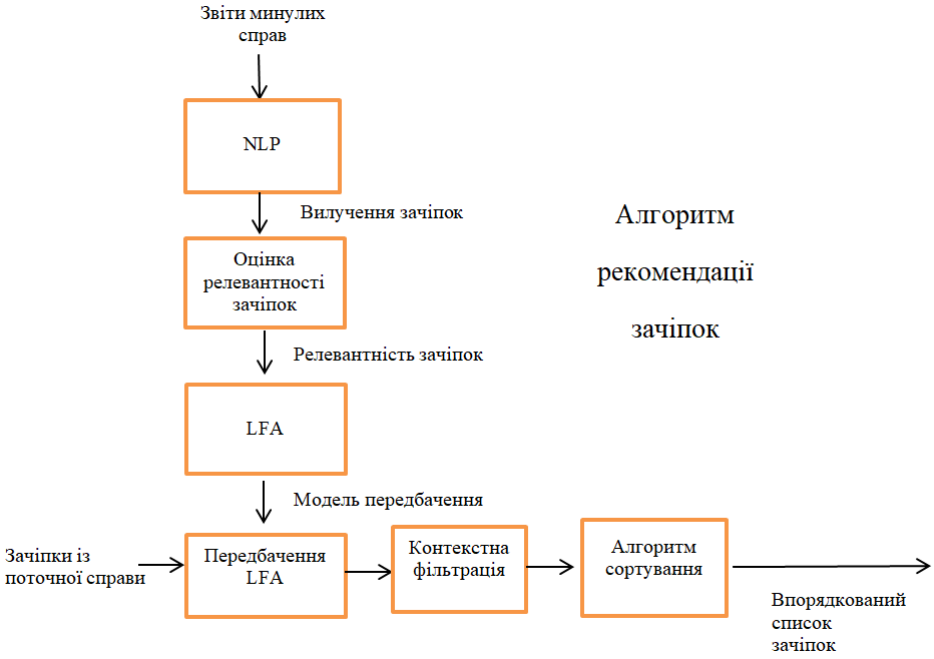

Рис. 2. Алгоритм рекомендації зачіпок

Перетворення від інформації про минулі справи до відповідних зачіпок залежить від типу наявного джерела. Якщо ми використовуємо текстові звіти, то нам необхідний NLP (*Natural Language Processing* – обробка природньої мови) модуль. Якщо у нас є зображення чи відео файли замість звітів, ми замінимо NLP модуль на модуль аналізу зображень чи відео. В основі NLP модуля, який може забезпечити важливі зачіпки, лежать добре відомі техніки, такі як: розпізнавання іменних об'єктів і виявлення предметів [3]. Застосування обох технік на звітах дозволяє вилучити основні зачіпки у вигляді пар ключ-значення (*key-value pairs*).

Релевантність зачіпок у кожній справі оцінюється у наступному модулі, який назвемо «Оцінка релевантності зачіпок». У ньому використаємо алгоритм релевантності tf-idf [4]. На цьому етапі інформація виражається так:

$$< c_m, l_n, r_{m,n} >, \quad (1)$$

де c_m – наявні m попередніх справ, l_n – n різних зачіпок вилучених із попередніх справ і $r_{m \cdot n}$ – виражає релевантність кожної із n зачіпок до m попередніх справ. Отже, релевантність зачіпок у справі виражена матрицею $R_{m,n}$. Цю матрицю можна розкласти на добуток матриць, описують властивості справ і зачіпок. Кожен рядок у $C_{m,p}$ – це вектор спорідненостей кожної справи і властивостей, кожен рядок у $I_{p,n}$ – це вектор, що виражає відношення між зачіпками і властивостями. Розклад матриці виконується у модулі LFA (*Latent Factor Analysis* – латентний факторний аналіз) у тих випадках, коли неможливо використати такі поширені техніки як метод Ланцоша для сингулярного розкладу матриці (SVD). Наш випадок саме такий, бо матриця релевантності є

розрідженою і частково визначеною та відсутні значення не можна інтерпретувати як нулі.

$$\begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{pmatrix} = \begin{pmatrix} c_{11} & \cdots & c_{1p} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mp} \end{pmatrix} \cdot \begin{pmatrix} l_{11} & \cdots & l_{1n} \\ \vdots & \ddots & \vdots \\ l_{p1} & \cdots & l_{pn} \end{pmatrix}. \quad (2)$$

Розклад повинен виконуватись тільки на відомих значеннях. Наше завдання – знайти множину векторів властивостей справ і зачіпок, які мінімізують квадратичну похибку на відомих значеннях:

$$\min_{U,M} \sum (r_{ij} - l_j c_i)^2 \quad (3)$$

Модель містить гіперпараметри, регуляризацію та швидкість навчання, які потрібно вибрати. Параметри регуляризації дуже важливі, оскільки важливо уникати перенавчання, що є єдиною проблемою такого роду даних.

Регуляризацію вводимо за допомогою параметра λ і норми вектора c_i і l_i. Крім того, параметр $t(i, j)$ використовується для налаштування точності кожного значення:

$$f(C, L) = \sum t(i, j)(r_{ij} - l_j c_i)^2 + \lambda \left(\sum \|c_i\|^2 + \sum \|l_i\|^2 \right). \quad (4)$$

Обрана методика для вирішення моделювання – альтернативний метод найменших квадратів (ALS – Alternating Least Squares), який є широко використовуваним алгоритмом у латентному факторному аналізі (LFA) алгоритмів колаборативної фільтрації. Ця модель використовується у модулі LFA для передбачення релевантності імовірної зачіпки у даній справі:

$\hat{r}_{c,t} = f(c, l)$ (5).

Список релевантних значень для найбільш вірогідних зачіпок присвоюється даній справі. Наступний модуль «Контекстна фільтрація» фільтрує результати, прибираючи зачіпки і змінюючи ваги їх значень в залежності від конкретних властивостей таких як місце або час даної справи. Ця операція виконується алгоритмом на основі правил. Після цього застосовується «Алгоритм сортування» для сортування зачіпок за їх релевантністю у даній справі і видача результатів слідчому.

Алгоритм рекомендації підозрюваних. В основі другого функціоналу лежить алгоритм колаборативної фільтрації на базі пам'яті, так як у цьому випадку LFA є некоректним для передбачень можливих підозрюваних. Властивості підозрюваних є дуже специфічними для кожної особи, тому шаблони можуть виявитись не підходящими для конкретної справи і алгоритм на основі пам'яті є більш ефективним. В основі алгоритму лежить така ідея: деякі із можливих підозрюваних у даній справі могли в минулому бути задіяні в інших схожих справах. Тому фільтрування найімовірніших відомих підозрюваних є дуже корисним для слідчих, яке заощадить багато часу за допомогою засобів імовірнісного зважування.

Крім того, коли слідчий має справу із новим злочином, де немає очевидних підозрюваних, імовірного підозрюваного можна взяти як зачіпку для вибору імовірних властивостей справді задіяних підозрюваних. Для цієї цілі використовують проміжну стадію KNN (*k nearest neighbors – k* найближчих сусідів), результатом якої стане зважене сусідство між справами.

Рис. 3 представляє процес «Алгоритму рекомендації підозрюваних», який має деякі спільні частини із алгоритмом рекомендації зачіпок.

Рис. 3. Алгоритм рекомендації підозрюваних

У даному функціоналі задіяні вектори визначаються так:

$$< c_m, s_n, r_{m \cdot n} > , \tag{6}$$

де c_m – це m наявних минулих справ, s_n – це n різних підозрюваних, вилучених із минулих справ, і $r_{m \cdot n}$ виражає задіяність кожного підозрюваного в попередніх справах.

Після вилучення властивостей за допомогою модуля NLP (чи аналогічних модулів для зображень, відео і аудіо), найважливішим кроком є обчислення сусідства поточної справи. Це виконується за допомогою KNN алгоритму, який обчислює K найбільш схожих справ до поточної справи, використовуючи модель, яка попередньо навчилася із подібностей і вдалих передбачень в минулих справах.

Використовуючи сусідство поточної справи, виконується звичайне колаборативне передбачення:

$$r_{c,s} = \bar{r}_s + \frac{\sum sim(c,c')(r_{s,c'} - \bar{r}_c')^2}{\sum |sim(c,c')|}, \tag{7}$$

де: $r_{c,s}$ – це імовірнісне передбачення зв'язку між імовірним підозрюваним s і поточною справою c, $\bar{r}_s$ виражає середнє залучення підозрюваних s у минулих справах, $sim(c,c')$ виражає схожість між поточною справою c і минулою справою c', і $\bar{r}_c'$ – це середнє залучення підозрюваних у справі c' (чим вище значення, тим більше підозрювані були задіяні у справі і, як наслідок, релевантність кожного із них має бути відносно менш важливою).

Модулі «Контекстне фільтрування» і «Алгоритм сортування» застосовуються, щоб отримати кінцевий список підозрюваних, аналогічно як і в першому функціоналі.

Рекомендаційні системи підходять для джерел неоднорідних даних, таких як ті, що

збирають під час кримінальних розслідувань. У цій статті була розглянута архітектура, яка застосовує зазначені техніки з двома цілями: надати рекомендацію найбільш релевантної одиниці серед зібраної інформації і запропонувати можливі напрями розслідування, покладаючись на докази і минулі справи. Контекстна інформація також використовуватиметься для визначення обмежень для аналізу цифрових даних.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. *Bobadilla J.* Recommender systems survey / J. Bobadilla, F. Ortega, A. Hernando [et. al.] // Knowledge-Based Systems. – **46**. – 2013. – P. 109-132.
2. *Quintana M.* Recommendation techniques in forensic data analysis: new approach / M. Quintana, S. Uribe, F. S?nchez [et. al.] //6th International Conference on Imaging for Crime Prevention and Detection (ICDP-15). – 2015. – P. 15-20.
3. *Temizer M*. Automatic Subject-Object-Verb relation extraction / Temizer, M., Diri, B. // International Symposium on Innovations in Intelligent Systems and Applications (INISTA). – 2012. – P. 244-248.
4. *Wu H*. C. Interpreting tf-idf term weights as making relevance decisions. / Wu H. C., Luk R. W. P., Wong K. F. [et. al.] // ACM Transactions on Information Systems (TOIS). – 2008. – Vol. 26, N. 3, Article 13. – P. 1-37.

RECOMMENDATION TECHNIQUES IN FORENSIC DATA ANALYSIS

Y. Eines

Ivan Franko National University of Lviv,
1 Universytetska Str., UA –79000 Lviv, Ukraine
eines.yuliia@gmail.com

Data mining for digital forensic analysis is focused on pattern extraction from large-scale data. These patterns are used to help analysts to solve crimes. One of the most promising applications of data mining algorithms is building recommendation systems. The goal of such systems is to propose future directions of the investigation. They are especially useful for investigation with large scale collections of forensic data.

In this paper I will depict one of possible architectures aiming to build recommender system for forensic data analysis.

This system consists of two subsystems: clues recommendation and suspects recommendation. They can be used together as well as independently.

The goal of the first subsystem is to make a list of clues based on previous investigations of other cases. First stage of clues recommendation algorithms is feature extraction. Feature extraction could be performed by NLP (Natural Language Processing) algorithms if we have text data or some other algorithms for analysis of images, videos and audios. Then for clues recommendation matrix decomposition LFA (Latent Factor Analysis) method is used. To make predictions also LFA is used. Recommendation is made by means of context filtering. The last stage is sorting of resulting data output of recommendation.

The goal of the second subsystem is to make a recommendation list of suspects. This is achieved through analysis of previous cases of the same type as current case. First stage of suspect recommendation algorithm is also feature extraction, which is performed the same way as in clues recommendation algorithm. Then KNN (K nearest neighbors) algorithms is used. Next stage is

collaborative prediction. Same as in clues recommendation algorithm recommendation is made by means of context filtering. The last stage is sorting of resulting data and output of recommendation.

Such system should help investigators to save time which is very important at first stage of investigation.

Key words: forensic analysis, recommendation, collaborative filtering.

Стаття: надійшла до редакції 20.11.2018,
доопрацьована 22.11.2018,
прийнята до друку 23.01.2018.