УДК 004.032.6

А. Б. Демчук
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

ВИКОРИСТАННЯ АСОЦІАТИВНИХ ПРАВИЛ ДЛЯ ВИРОБЛЕННЯ ЗНАНЬ З ПОБУДОВИ ТИФЛОКОМЕНТАРІВ

Описано розроблення математичного забезпечення процесу тифлокоментування відеоконтенту за асоціативними правилами. Це дало змогу формалізувати побудову відеоконтенту для осіб з вадами зору.

Ключові слова: тифлокоментування, аудіодескрипція, асоціативні правила, відеоконтент, інформаційні технології, відеоконтент для осіб з вадами зору.

The development of mathematical support process of typhlocomment video content through the use of of associative rules is discribed. This made it possible to formalize the construction of video content for people with visual impairments.

Key words: typhlocomment, audiodescription, association rules, videocontent, IT, videocontent for sightless.

Вступ

Алгоритми пошуку асоціативних правил стали зараз одним з популярних методів виявлення прихованих закономірностей та побудови знань. Перший алгоритм пошуку асоціативних правил – так званий AIS – розробили 1993 року співробітники дослідницького центру IBM Almaden. Після цієї початкової роботи зросло зацікавлення асоціативними правилами; на середину 90-х років минулого століття припав пік дослідницьких робіт у цій області, і відтоді щороку з’являлося по кілька алгоритмів [1]. Уперше задачу пошуку асоціативних правил було запропоновано для визначення типової поведінки покупців під час купівлі у супермаркетах, тому іноді її ще називають аналізом ринкового кошика (market basket analysis). Реєструючи всі бізнес-операції протягом певного часу своєї діяльності, торгові компанії накопичують величезну кількість таких записів про придбання. Кожний такий запис називається транзакцією. Момент входження покупця в магазин вважається моментом початку, або відкриття транзакції. Момент оплати покупцем вибраних товарів вважається закриттям транзакції. У базі даних зберігаються записи тільки про закриті транзакції (успішні купівлі). Якщо покупець вийшов з магазину, нічого не придбавши, то транзакція вважається не закритою, і в базу даних нічого не записується.

Загальна постановка проблеми

Описати використання асоціативних правил для вироблення знань з побудови тифлокоментарів.

Формулювання мети

Описано використання асоціативних правил для вироблення знань з побудови тифлокоментарів. Обрано алгоритм Apriori, як оптимальний для поставленої задачі.

Аналіз наукових результатів

Одна з найпоширеніших задач, що розв’язуються під час аналізу баз даних, полягає у пошуку закономірностей або наборів закономірностей, що одночасно зустрічаються в багатьох правилах. У нашому випадку це питання, на які було дано ствердні відповіді [2–4].

Нехай I – скінченна множина елементів. Нехай D – набір рядків інформаційної таблиці. Кожному такому рядку відповідає транзакція T, що складається з підмножини елементів I, $T \subseteq I$.

Елемент входить в транзакцію, якщо у відповідному рядку відповідний атрибут набуває значення „1”. Ми говоримо, що транзакція T містить X, деяку підмножину елементів I, якщо $X \subseteq T$. Асоціативним правилом називається імплікація $X \Rightarrow Y$, де $X \subset I$, $Y \subset I$ та $X \cap Y = \varnothing$.

Кажуть, що правило $X \Rightarrow Y$ має підтримку (support) s, якщо s% транзакцій з D, містять множину $X \cup Y$,

$$\sup p(X \Rightarrow Y) = \sup p(X \cup Y) = \frac{count(T : X \cup Y \subseteq T)}{size(D)} \cdot 100\ \% \qquad (1)$$

Достовірність (confidence) правила показує, яка ймовірність того, що з X випливає Y. Кажуть, що правило $X \Rightarrow Y$ справедливе з достовірністю c, якщо c% транзакцій з D, що містять X, також містять Y,

$$conf(X \Rightarrow Y) = \frac{\sup p(X \cup Y)}{\sup p(X)} \cdot 100\ \%. \qquad (2)$$

У нашому випадку загальна кількість транзакцій – 100.

Множина {Чи запам'ятали ви головних героїв відеоконтенту?} зустрічається 93 рази, а , отже, значення *support* множини {Чи запам'ятали ви головних героїв відеоконтенту?} дорівнюватиме:

$$supp(\{\text{Чи запам'ятали ви головних героїв відеоконтенту?}\}) = \frac{93}{100} \cdot 100\ \% = 93\ \%.$$

Аналогічно:

supp({Чи зрозуміли ви побут (місцевість, предмети) героїв, які були важливі для сприйняття відеоконтенту?, Чи задоволені ви переглядом цього відеоконтенту з тифлокоментарем?}) = 95 %;

supp({Чи запам'ятали Ви назву відеоконтенту? Чи пригадаєте назву студії, яка зняла цей відеоконтент? Чи запам'ятали ви головних героїв відеоконтенту?}) = 70 %;

supp({Чи не було моментів у відеоконтенті, коли Ви перестали розуміти те, що відбувається на екрані(незрозумілі звуки, які не були описані тифлокоментарем)? Чи не було моментів, коли слова тифлокоментатора накладались на слова акторів, важливі звуки, і тим заважали сприйняттю відеоконтенту? Чи не був тифлокоментар заплутаний, незрозумілий? Чи були терміни, які ви не зрозуміли? Чи не був тифлокоментар надмірним/недостатнім?}) = 81%;

Обчислимо достовірність правила {Чи задоволені ви переглядом цього відеоконтенту з тифлокоментарем?} $\Rightarrow$ {Чи зрозуміли ви сам сюжет, його розвиток протягом відеоконтенту, розв'язку?}

$$\text{conf()} = \frac{95\ \%}{100\ \%} \cdot 100\ \% = 95\ \%.$$

Отже, метою пошуку асоціативних правил є встановлення таких залежностей: якщо зустрівся деякий набір елементів X, то на підставі цього можна зробити висновок про те, що інший набір елементів Y також повинен зустрітися з певною імовірністю.

Пошук асоціативних правил – зовсім не тривіальна задача, як може видатися на перший погляд. Одна з проблем – алгоритмічна складність при знаходженні наборів елементів, що часто зустрічаються (frequent itemsets), тому що з ростом кількості елементів у множині I експоненційно росте кількість потенційних frequent itemsets, хоча реальна кількість frequent itemsets може виявитися набагато меншою. Наприклад, якщо множина I містить 3 елементи, то кількість потенційних frequent itemsets дорівнюватиме 7, при 4 елементах – 15, при 5 – 31 тощо. Кількість потенційних frequent itemsets обчислюють за формулою:

$$N = 2^{|I|} - 1 \qquad (3)$$

Алгоритми пошуку асоціативних правил призначені для знаходження правил $X \Rightarrow Y$, причому значення support і confidence цих правил мають бути вищі від деяких наперед визначених граничних значень, що називаються, відповідно, мінімальною підтримкою (minsupport) і мінімальною достовірністю (minconfidence).

Значення параметрів minsupport і minconfidence вибирають так, щоб обмежити кількість знайдених правил. Якщо minsupport має велике значення, то алгоритми знаходитимуть правила, добре відомі аналітикам або настільки очевидні, що немає сенсу здійснювати такий аналіз. З іншого боку, низьке значення minsupport веде до ґенерування величезної кількості правил, що зазвичай вимагає істотних обчислювальних ресурсів. Проте, більшість цікавих правил знаходять саме при низькому значенні minsupport. Хоча занадто низьке значення minsupport веде до ґенерування статистично не обґрунтованих правил.

У нас використано 10 різних питань, тому існує 2^{10}–1=1023 потенційних frequent itemsets. Задачу знаходження асоціативних правил поділимо на дві підзадачі:

1) знаходження всіх frequent itemsets;

2) формулювання правил зі знайдених frequent itemsets.

Один з перших алгоритмів, що ефективно розв'язують подібний клас задач, – це алгоритм Apriori (див. рис. 1, 2).

Крок 1. Обчислення значення *support* усіх 1-елементних наборів та відкидання тих, що не є *frequent.*
Крок 2. Ґенерування кандидатів.
Крок 3. Підрахунок значення *support* для отриманих кандидатів та відкидання тих, що не є *frequent.*
Крок 4. Перевірка умов зупинки (досягнення певного розміру наборів або випадок, коли після 3-го кроку залишилося $n \leq 1$ *frequent itemsets*) і, в разі виконання умов – перехід на 5-й крок, інакше – на 2-й крок.
Крок 5. Генерування правил.

Рис. 1. Кроки алгоритму Apriori

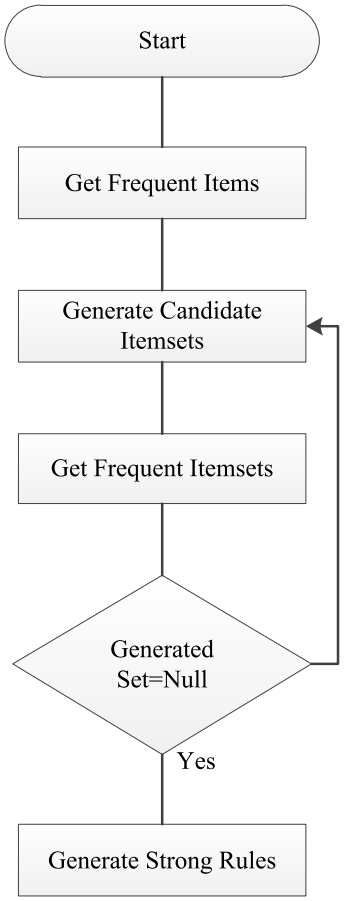

Рис. 2. Блок-схема алгоритму Apriori

Розглянемо кроки алгоритму детальніше.

На першому кроці алгоритму Apriori знаходяться 1-елементні *frequent* набори. Для цього необхідно просканувати всю інформаційну таблицю, підрахувати значення *support* для цих наборів та відкинути ті, для яких це значення менше за значення *minsupport.*

На другому кроці для генерування кандидатів немає необхідності знову звертатися до бази даних. Щоб одержати *k-itemsets*, скористаємося *(k-1)-itemsets*, що є *frequent*. Ґенерування кандидатів складатиметься з двох частин.

1. *Об'єднання*. Кожен кандидат формуватиметься розширенням *frequent (k-1)-itemset* додаванням елемента з іншого *frequent (k-1)-itemset*. Алгоритм генерування *k-itemsets* можна записати покроково:

Крок 1. Вибираємо *frequent (k-1)-itemset.*

Крок 2. Об'єднуємо його по черзі з останніми елементами всіх інших *frequent (k-1)-itemsets*, в яких останні елементи є лексикографічно більшими за останній елемент вибраного на першому кроці *frequent (k-1)-itemset.*

Крок 3. Повторюємо крок 1 та крок 2 для всіх *frequent (k-1)-itemsets.*

2. *Видалення надлишкових правил.* На підставі властивості *антимонотонності* видаляються всі утворені *k-itemsets*, хоча б одна з (*k-1*)-елементних підмножин яких не є *frequent*.

На третьому кроці алгоритму, після ґенерування кандидатів, наступною задачею є підрахунок значення *support* для кожного кандидата. Для цього, як і на першому кроці, сканують інформаційну таблицю. Після підрахунку ті кандидати, для яких значення *support* є меншим за встановлене значення *minsupport*, – видаляються.

На п'ятому кроці алгоритму, після того як знайдені всі *frequent itemsets*, відбувається ґенерування правил. Ґенерування правил – менш трудомістка задача, ніж знаходження всіх *frequent itemsets*. По-перше, для підрахунку достовірності правила достатньо знати значення *support* самого правила і множини, що є головою правила.

Знайдені правила записані в онтологію нозологій у вигляді SWRL-правил. Приклад правила отримання адаптованого відеоконтенту:

$$Adapting\_videocontent(?c) \leftarrow Computer(?c) \wedge hasAudioEditor(?c,?ae) \wedge$$
$$hasVideoContent(?ae,?vc) \wedge needForAdaptation(?vc)$$

Щоб отримати знання у вигляді правил побудови тифлокоментарів, проведено опитування людей з вадами зору у вигляді анкети. Після перегляду відеоконтенту з тифлокоментарями було задано такі 10 питань (таблиця).

Запитання, поставлені після перегляду відеоконтенту, адаптованого для незрячих та осіб з вадами зору

1. Чи запам'ятали Ви назву відеоконтенту?
2. Чи пригадаєте назву студії, яка зняла цей відеоконтент?
3. Чи запам'ятали ви головних героїв відеоконтенту?
4. Чи зрозуміли ви сам сюжет, його розвиток протягом відеоконтенту, розв'язку?
5. Чи не було моментів у відеоконтенті коли Ви перестали розуміти те, що відбувається на екрані (незрозумілі звуки, які не було описано тифлокоментарем)?
6. Чи не було моментів, коли слова тифлокоментатора накладались на слова акторів, важливі звуки, і тим самим заважали сприймати відеоконтент?
7. Чи зрозуміли ви побут (місцевість, предмети) героїв, які були важливі для сприйняття відеоконтенту?
8. Чи не був тифлокоментар заплутаний, незрозумілий? Чи були терміни, яких ви не зрозуміли?
9. Чи не був тифлокоментар надмірним/недостатнім?
10. Чи задоволені ви переглядом цього відеоконтенту з тифлокоментарем?

Відповіді на питання давались у вигляді 1-Так, 0-Ні. Отримані результати відповідей наведено на рис. 3. Відповідні взаємодії поставлених запитань наведено на рис. 4. Для опрацювання отриманих анкет використано теорію асоціативних правил.

	№ запитання										%
№ опитаного	1	2	3	4	5	6	7	8	9	10	задоволення
1	1	1	1	1	1	1	1	1	0	1	90,00%
2	1	1	1	1	1	1	1	1	1	1	100,00%
3	1	0	1	1	1	1	1	1	1	1	90,00%
4	1	1	1	1	1	1	1	0	1	1	90,00%
5	1	1	1	1	1	1	0	1	1	1	90,00%
.			.	.	.						
98	1	0	1	1	1	1	1	1	1	1	90,00%
99	1	0	1	1	1	1	1	1	1	1	90,00%
100	1	1	1	1	1	0	1	1	1	1	90,00%
% відповідей	100,0%	71,%	93,%	95,%	96,%	94,%	95,%	94,%	97,%	100,0%	
											AVERAGE
											93,50%

групи запитань	average
1,2,3	88,0%
4,7	95,0%
5,6,8,9	95,3%
10	100,0%

Рис. 3. Результати відповідей респондентів та середнє значення за групами запитань

Дія1	Дія2	Дія3	Дія4	Дія5
7and10	1and2and3	4or7	5and6and8and9	10=>4
95,0%	70,0%	98,0%	81,0%	95,0%

Рис. 4. Результати взаємодій поставлених запитань

Висновки та перспективи подальших наукових розвідок

Описано використання асоціативних правил для вироблення знань з побудови тифлокоментарів, що дало змогу знайти закономірності між зв'язаними подіями. Обрано алгоритм Apriori як оптимальний для цієї задачі. Реалізовано застосування асоціативних правил для створення взаємодій за результатами роботи програмного комплексу "Audio Editor", який розроблений для вирішення задачі адаптації відеоконтенту для осіб з вадами зору.

1. Буров К. Обнаружение знаний в хранилищах данных // Открытые системы. – 1999. – № 5–6. – С. 67–77. 2. Литвин В. В. Проблема автоматизованої розбудови базової онтології / В. В. Литвин, Т. М. Черна // Вісник Нац. ун-ту "Львівська політехніка" – Львів, 2014. – № 805. – С. 306–315. 3. Дюк В., Самойленко А. Data Mining: Учебный курс. – СПб: Питер, 2001. – 368 с. 4. Han J., Kamber M. Data Mining: Concepts and Techniques. – Simon Fraser University, 2000.