УДК 519.6

ЗАСТОСУВАННЯ МЕТОДІВ DATA MINING ДЛЯ АВТОМАТИЧНОГО ФОРМУВАННЯ РЕКОМЕНДАЦІЙ

Чертов О.Р., Александрова М.В.

DATA MINING METHODS USAGE FOR THE TASK OF AUTOMATED FORMING OF RECOMMENDATIONS

Chertov O., Aleksandrova M.

В статті запропонований новий алгоритм автоматичного формування рекомендацій, що використовує техніки інтелектуального аналізу даних в якості окремого етапу. Також наведено експериментальний приклад застосування запропонованого алгоритму.

Ключові слова: *системи рекомендацій, інтелектуальний аналіз даних, пошук асоціативних правил.*

Вступ. Системи рекомендацій

Системи рекомендацій — це програмне забезпечення та набір технологій, метою роботи яких є спроба визначити об'єкти, які можуть зацікавити користувача [1]. Рекомендації можуть мати як комерційний (рекомендації товарів на сайтах електронної комерції, наприклад, Інтернет-магазин Amazon.com), так і некомерційний характер (рекомендації відеозаписів з безкоштовним переглядом, наприклад, сайт YouTube).

В переважній більшості випадків рекомендаційні системи формують персоналізовані рекомендації у вигляді переліку рекомендованих об'єктів. Останні ранжуються в порядку зменшення цільової функції, яка представляє собою очікувану корисність об'єкту для активного користувача, тобто користувача, для якого формується рекомендація. Для обчислення конкретних значень цільової функції використовується інформація про вподобання користувачів, які вони виражають в різних формах: явній (наприклад, встановлюючи рейтинги для об'єктів рекомендацій) або неявній (користувач здійснив електронну покупку рекомендованого об'єкту, отже, останній був для нього корисним).

Ідеологія рекомендаційних систем базується на дуже простому правилі: в повсякденному житті люди орієнтуються на рекомендації, які були зроблені тими, кого вони вважають експертами в даній галузі [2]. Наприклад, при виборі книжки людина часто користується порадами своїх друзів зі схожими літературними вподобаннями; роботодавці, наймаючи на роботу нового працівника, враховують інформацію, надану в рекомендаційних листах; обираючи фільм для перегляду, людина, зазвичай, перечитує відгуки кінокритиків.

Намагаючись використовувати це правило, більшість рекомендаційних систем пропонують активному користувачеві такі об'єкти, які були високо оцінені іншими користувачами зі схожими уподобаннями. Такий підхід називається колаборативною фільтрацією [3] та базується на припущенні що, якщо в минулому активний користувач мав схожі уподобання з деякими іншими користувачами, то так буде і надалі.

Алгоритм пошуку важелів впливу

В роботах [4-6] авторами були розроблені декілька версій алгоритму пошуку важелів впливу. Задачею цього алгоритму є визначення факторів, за допомогою яких можна стимулювати в певному напрямку процес формування людиною відповіді на такі питання як:

- Чи переїздити до іншого міста?
- Чи починати навчання?
- Чи народжувати дитину?

Алгоритм було протестовано на п'ятивідсотковій вибірці даних перепису населення (США, штат Каліфорнія, 2000 р.) [7]. Метою експериментів було визначення факторів підвищення народжуваності.

В загальному випадку запропонований алгоритм допомагає виділити ті характеристики, які необхідно надати деякому об'єкту для того, щоб він перемістився з однієї контрастної множини до іншої. Контрастними множинами в даній ситуації називаються декілька підмножин початкового набору даних, елементи яких однозначно визначаються значеннями деяких атрибутів (параметрів контрастності) і не можуть одночасно належати двом або більше контрастним множинам, наприклад:

- кількість дітей в сім'ї: 2 або більше, 1 дитина, сім'ї без дітей (3 контрастні множини),
- стать: чоловік, жінка (2 контрастні множини).

Аналізом контрастних множин називають процес виділення шаблонів та закономірностей, що відрізняють їх між собою [8].

Головною особливістю запропонованого алгоритму є використання технік Data Mining в якості окремого кроку аналізу. В роботах [4, 5] було застосовано техніку кластеризації, що надало можливість отримати набір правил вигляду:

- молодим сім'ям з низьким рівнем освіти необхідно надати дешеве орендоване житло;
- сім'ям середньої вікової групи необхідно надати дешеві кредити для придбання власного житла.

Застосування цих рекомендацій дозволяє наблизитися до поставленої мети (в нашому випадку — підвищити рівень народжуваності). В роботі [6] замість кластеризації, був застосований алгоритм пошуку асоціативних правил, в результаті чого були сформовані більш точні рекомендації. Наприклад, аналізуючи пару контрарних правил вигляду

- IF <wife work class=employee of private for profit company> AND <detached house=yes> THEN <children=yes> ($supp = 21{,}5\%$, $conf = 72{,}7\%$)
- IF <wife work class=employee of private for profit company> AND <detached house=no> THEN <children=no> ($supp = 19{,}5\%$, $conf = 72{,}5\%$)

можна сказати, що якщо тим сім'ям, де дружина є працівником приватної комерційної організації, надати відокремлений будинок, то з високою ймовірністю (72,7%) подружжя народить дитину. Також з характеристик правил видно, що така рекомендація може бути застосована до 19,5% сімейних пар. Парою контрарних правил в даному випадку називаються два правила, у висновку яких містяться параметри контрастності з різними значеннями.

Постановка завдання

Метою даної роботи є адаптація алгоритму пошуку важелів впливу до задачі формування автоматичних рекомендацій деяких об'єктів (книжок, фільмів, товарів тощо).

Адаптація алгоритму пошуку важелів впливу до задачі формування рекомендацій

Алгоритм пошуку важелів впливу був вперше запропонований в роботі [4]. Він базується на застосуванні техніки жорсткої кластеризації (subtractive clustering [9]) і складається з наступних кроків:

1. Виділити із початкової множини записів про респондентів дві контрастні групи N_1 та N_2. Перша група повинна містити записи про тих респондентів, які володіють характеристикою, на наявність якої потрібно впливати; друга, навпаки, повинна містити записи про тих, хто не володіє нею. Також на виділені групи можуть накладатися додаткові обмеження, обумовлені проблемною областю.
2. Визначити атрибути, які можуть потенційно впливати на наявність обраної характеристики. Виділити атрибути для кластеризації, тобто такі, що є числовими, або можуть бути порівняні за допомогою чисел. Виходячи із специфіки поставленої задачі визначити інваріантні параметри для обох груп, тобто такі параметри, на значення яких неможливо або дуже важко впливати зовнішньо, наприклад, вік, стать, етнічне походження тощо.
3. Провести кластеризацію групи N_1, розділити її на підгрупи.
4. Визначити границі кожного з інваріантних параметрів в межах підгруп.
5. Використовуючи отримані в п. 4 межі, виділити з групи N_2 прототипи підгруп з N_1.
6. Порівняти характеристики отриманих під час кластеризації підгруп та їх прототипів.

В статті [5] результати роботи запропонованого алгоритму були покращені за рахунок використання алгоритму нечіткої кластеризації (fuzzy c-means clustering [10]). В роботі [6] замість кластеризації був використаний алгоритм пошуку асоціативних правил, кроки алгоритму пошуку важелів впливу були змінені (рис. 1).

Для адаптації алгоритму пошуку важелів впливу до області систем рекомендацій необхідно визначити наступні параметри алгоритму:

1) параметр контрастності,
2) параметри, що потенційно можуть впливати на значення параметру контрастності,
3) інваріантні параметри.

Найбільш доцільним видається адаптація останньої версії алгоритму (на базі асоціативних правил), оскільки вона надає найбільш персоналізовані рекомендації.

Експериментальні результати

Для проведення експерименту було взято множину даних Anonymous Microsoft Web Data [11], яка використовувалась для тестування алгоритмів автоматичного формування рекомендацій [12]. Дані представляють собою історію відвідування сайту www.microsoft.com впродовж одного тижня в лютому 1998 р. та містять записи про послідовність відвідування сторінок сайту 38 тисячами анонімними користувачами.

Метою експерименту в роботі [12] було прогнозування за допомогою алгоритмів колаборативної фільтрації сторінок сайту, які користувач відвідає, на базі аналізу попередньо відвіданих сторінок.

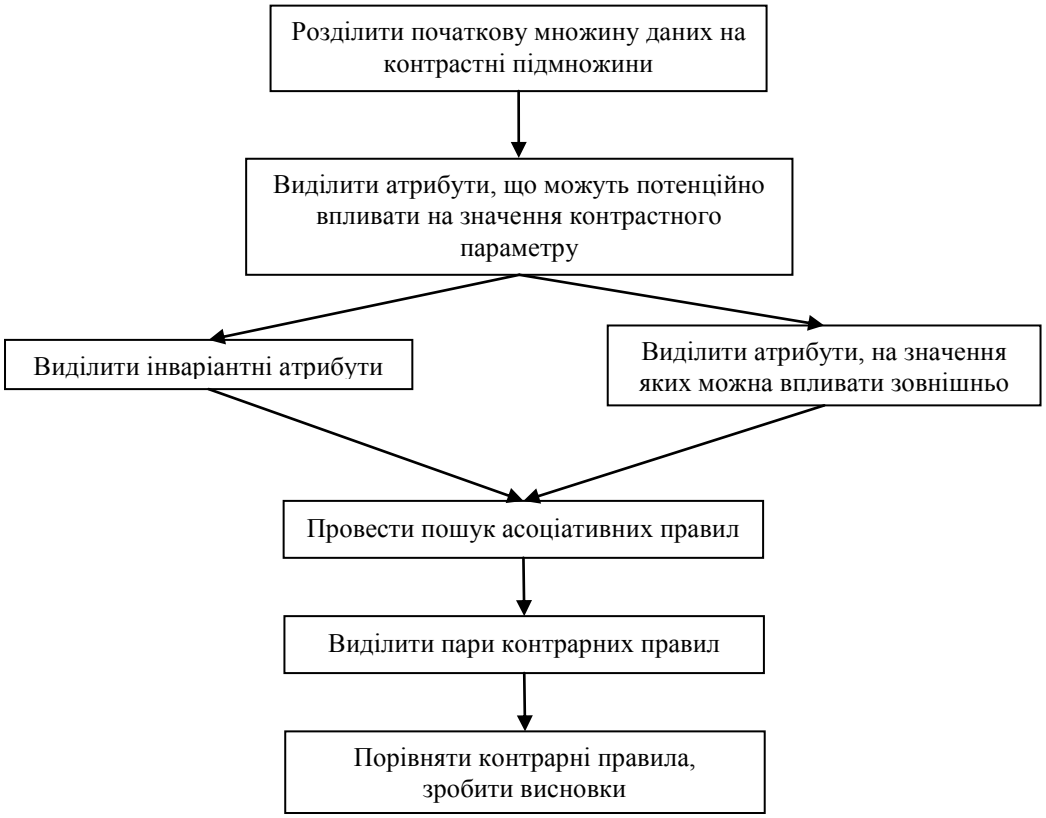

Рис. 1. Схема алгоритму пошуку важелів впливу на базі техніки виділення асоціативних правил

Припустимо, що у компанії є деякі стратегічні інтереси, і вона хоче підвищити відвідуваність визначених сторінок свого сайту. Для досягнення цієї мети на поточній відвідуваній сторінці можна автоматично генерувати посилання на інші сторінки сайту таким чином, щоб привести користувача на одну з бажаних сторінок. Користувач може перейти за одним із запропонованих посилань або ні. Необхідно визначити, які сторінки потрібно рекомендувати користувачу, щоб він відвідав одну з цільових сторінок.

Для розв'язання поставленої вище задачі можна використовувати алгоритм пошуку важелів впливу. Параметром контрастності в даному випадку буде факт відвідування або невідвідування цільової сторінки. В якості параметрів, що можуть впливати на значення параметру контрастності, будуть використовуватися попередньо відвідані впродовж сесії сторінки. Інваріантним параметром буде сторінка входу на сайт, оскільки на її визначення стороння система не може впливати (вона обирається без рекомендування).

В множині даних Anonymous Microsoft Web Data зафіксована інформація про відвідування 294 сторінок сайту. В якості цільових в рамках даного експерименту були обрані такі, що містять інформацію про підтримку продуктів Microsoft або якимось чином пов'язані із фінансовими потоками. Обрані 36 сторінок наведені в табл. 1.

Таблиця 1

Сторінки, які були обрані в якості цільових в рамках експерименту

ID	NAME	ID	NAME	ID	NAME
1121	Microsoft Shop	1162	IIS Support	1013	Visual Basic Support
1145	Visual FoxPro Support	1046	IE Support	1249	Fortran Support
1276	Visual Test Support	1197	SQL Support	1192	Visual J++ Support
1147	Microsoft Financial Forum	1231	Windows NT Developer Support	1206	Volume Purchasing Options
1218	MS Publisher Support	1214	MS Financial Services	1230	Mail Support
1205	Hardware Supprt	1049	Support Network Program Information	1035	Windows95 Support
1133	FrontPage Support	1226	MS Schedule+ Support	1077	MS Office Support
1132	MS Money Support	1184	MS Excel Support	1151	MS PowerPoint Support
1001	Support Desktop	1160	Visual C Support	1085	Exchange Support
1138	Developer Shop	1220	Mac Office Support	1090	Games Support
1211	SMSMGT Support	1168	Sales Information	1135	MS Word Support
1181	Kids Support	1161	Works Support	1210	SNA Support

Для проведення подальшого аналізу масив даних Anonymous Microsoft Web Data було модифіковано, зокрема,

- були обрізані всі послідовності відвідувань сторінок таким чином, щоб цільова сторінка (якщо вона присутня в послідовності), була останньою;
- з отриманих послідовностей були видалені всі, що містять менше трьох елементів, оскільки в такому випадку система може рекомендувати лише одну з цільових сторінок;
- масив даних був представлений у вигляді рядків, кожен з яких представляє собою послідовність відвідуваних сторінок одним користувачем.

Для пошуку асоціативних правил використовувалась система STATISTICA 10, в рамках якої реалізований алгоритм *Apriori* [13, 14]. Оскільки цей алгоритм не зважає на послідовність згадування елементу в транзакції (в нашому випадку елемент — це назва сторінки в одному рядку вхідних даних) і розрізняє елементи лише за їх написанням, було окремо виділено сторінку входу (*START=pageNM*). В кінці кожного запису було поставлено позначку про факт досягнення цільової сторінки (*END=YES*, *END=NO*) — див. рис. 2.

	Start page		**pages**	
user 1	*START=page11*	*page12*	*page13*	*END=YES*
user 2	*START=page21*	*page22*	*END=NO*	
…	*START=…*	…	…	…
user *N*	*START=pageN1*	*pageN2*	*pageN3*	…

END

Рис. 2. Формат вхідних даних для пошуку асоціативних правил

Оскільки сторінка входу є інваріантним параметром, який однозначно виділяється в масиві даних, то нас будуть цікавити лише такі пари контрарних правил, які або не мають фіксованої сторінки входу, або в яких вона є однаковою. Тому для прискорення роботи алгоритму пошуку асоціативних правил вхідну множину даних можна розділити на декілька підмножин за значенням цільової сторінки.

В рамках даного експерименту початкова множина даних була розділена на 5 підмножин таким чином: сторінка входу поточного елементу (послідовності відвіданих сторінок) використовується як вхідна більше 3000 разів (top3000), більше 1000 разів (top1000), більше 500 разів (top 500) або більше 100 разів (top100). Решта елементів була віднесена до підмножини Other. Кількість елементів в отриманих підмножинах склала 7296, 2753, 1380, 4640 та 2079 відповідно. В якості контрарних правил шукались такі, що мають необов'язково велику підтримку (min_support = 1%, для підмножини top3000 min_supp = 0,5%), високу достовірність (min_confidence = 60%) та висновками яких є «цільова сторінка досягнута» або «цільова сторінка не досягнута».

Контрарні правила були успішно сформовані для підмножин top1000 і Other, нижче наведені деякі з них:

1. Підмножина top1000
   1.1. IF <Start=Products> & <Microsoft_com_Search> & <Free_Downloads> THEN <END=YES> (supp=1,34%, conf=67,27%)
   1.2. IF <Start=Products> & <Microsoft_com_Search> THEN <END=NO> (supp=8,03%, conf=76,21%)
2. Підмножина Other
   2.1. IF <isapi> & <Microsoft_com_Search> & <Free_Downloads> THEN <END=YES> (supp=1,06%, conf=69,86%)
   2.2. IF <isapi> & <Microsoft_com_Search> THEN <END=NO> (supp=3,99%, conf=67,48%).

Аналізуючи отримані правила можна сказати, що для досягнення бажаної мети потрібно зробити наступне: якщо користувач зайшов на сайт через сторінку Products, а також відвідав сторінку Microsoft_com_Search, йому потрібно запропонувати відвідати сторінку Free_Downloads; в другому з наведених випадків, користувачеві, який зайшов через довільну сторінку, але також відвідав сторінки isapi і Microsoft_com_Search, також потрібно запропонувати переглянути сторінку Free_Downloads.

Висновки

В роботі описаний принцип адаптації алгоритму пошуку важелів впливу до задачі формування автоматичних рекомендацій. На експериментальному прикладі показано, що його застосування може бути успішним. Головною особливістю запропонованого алгоритму є те, що він може застосовуватися не тільки для надання рекомендацій, але також для визначення шляху, за яким користувача можна привести до певної бажаної цілі. Перспективами подальших досліджень є розробка версії алгоритму, що як контрастний параметр буде використовувати

рівень задоволеності користувача роботою системи.

Література

1. Ricci F. Introduction to Recommender Systems Handbook / F. Ricci, L. Rokach, B. Shapira // Recommender Systems Handbook, Berlin : Springer — 2011. — Chapter 1. — P. 1—35.
2. Mahmood T. Improving recommender systems with adaptive conversational strategies / T. Mahmood, F. Ricci // Proceedings of the 20th ACM conference on Hypertext and hypermedia, New York, USA. — 2009. — P. 73—82.
3. Breese J. Empirical analysis of predictive algorithms for collaborative filtering / J. Breese, D. Heckerman, C. Kadie // Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, San Francisco : Morgan Kaufmann Publishers Inc. — 1998. — P. 43—52.
4. Chertov O. Clustering with prototype extraction for census data analysis / O. Chertov, M. Aleksandrova // Proceedings of the World Conference on Soft Computing, WConSC-2011, San Francisco. — 2011. Available: http://arxiv.org/abs/1106.5122
5. Chertov O. Fuzzy clustering with prototype extraction for census data analysis / O. Chertov, M. Aleksandrova // Soft Computing: State of the Art Theory and Novel Applications.

Studies in Fuzziness and Soft Computing. — 2013. — Vol. 291. — P. 289—313.
6. Chertov O.Using Association Rules for Searching Levers of Influence in Census Data / O. Chertov, M. Aleksandrova // Procedia — Social and Behavioral Sciences. — 2013. — Vol. 73. — P. 475—478.
7. Minnesota Population Center, University of Minnesota. Integrated Public Use Microdata Series International. [Online]. Available: https://international.ipums.org/international/
8. Dong G. International Workshop on Contrast Data Mining and Applications. 2011. [Online]. Available: http://www.cs.wright.edu/~gdong/ContrastDMWorkshop.pdf
9. Generation of fuzzy rules with subtractive clustering / A. Priyono, M. Ridwan, A. Jais Alias et al. // Jurnal Teknology. — 2005. — Vol. 43. — P. 143—153.
10. Bezdek J.C. FCM: The fuzzy *c*-means clustering algorithm / J.C. Bezdek, R. Ehrlich, W. Full //," Computers & Geoscience. — 1984. — Vol. 10. — Is. 2–3. — P. 191—203.
11. Asuncion A., Newman D.J. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. [Online]. — 2007. — Available: http://www.ics.uci.edu/~mlearn/MLRepository.html
12. Breese J. Empirical Analysis of Predictive Algorithms for Collaborative Filtering / J. Breese, D. Heckerman., C. Kadie // Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison,, 1998. — 21 p.
13. STATISTICA Help. Sequence, Association, & Link Analysis (SAL) Technical Notes. [Online]. Available: http://documentation.statsoft.com/STATISTICAHelp.aspx?path=LinkAnalysis/Overviews/TechnicalNotes
14. Agrawal R. Fast algorithms for mining association rules / R Agrawal, R Srikant // Proceedings of the 20th International Conference on Very Large Data Bases, San Francisco. — 1994. — P. 487—499.

References

15. Ricci F. Introduction to Recommender Systems Handbook / F. Ricci, L. Rokach, B. Shapira // Recommender Systems Handbook, Berlin : Springer — 2011. — Chapter 1. — P. 1—35.
16. Mahmood T. Improving recommender systems with adaptive conversational strategies / T. Mahmood, F. Ricci // Proceedings of the 20th ACM conference on Hypertext and hypermedia, New York, USA. — 2009. — P. 73—82.
17. Breese J. Empirical analysis of predictive algorithms for collaborative filtering / J. Breese, D. Heckerman, C. Kadie // Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, San Francisco : Morgan Kaufmann Publishers Inc. — 1998. — P. 43—52.
18. Chertov O. Clustering with prototype extraction for census data analysis / O. Chertov, M. Aleksandrova // Proceedings of the World Conference on Soft Computing, WConSC-2011, San Francisco. — 2011. Available: http://arxiv.org/abs/1106.5122
19. Chertov O. Fuzzy clustering with prototype extraction for census data analysis / O. Chertov, M. Aleksandrova // Soft Computing: State of the Art Theory and Novel Applications.

Studies in Fuzziness and Soft Computing. — 2013. — Vol. 291. — P. 289—313.
20. Chertov O.Using Association Rules for Searching Levers of Influence in Census Data / O. Chertov, M. Aleksandrova // Procedia — Social and Behavioral Sciences. — 2013. — Vol. 73. — P. 475—478.
21. Minnesota Population Center, University of Minnesota. Integrated Public Use Microdata Series International. [Online]. Available: https://international.ipums.org/international/
22. Dong G. International Workshop on Contrast Data Mining and Applications. 2011. [Online]. Available: http://www.cs.wright.edu/~gdong/ContrastDMWorkshop.pdf
23. Generation of fuzzy rules with subtractive clustering / A. Priyono, M. Ridwan, A. Jais Alias et al. // Jurnal Teknology. — 2005. — Vol. 43. — P. 143—153.
24. Bezdek J.C. FCM: The fuzzy *c*-means clustering algorithm / J.C. Bezdek, R. Ehrlich, W. Full //," Computers & Geoscience. — 1984. — Vol. 10. — Is. 2–3. — P. 191—203.
25. Asuncion A., Newman D.J. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. [Online]. — 2007. — Available: http://www.ics.uci.edu/~mlearn/MLRepository.html
26. Breese J. Empirical Analysis of Predictive Algorithms for Collaborative Filtering / J. Breese, D. Heckerman., C. Kadie // Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison,, 1998. — 21 p.
27. STATISTICA Help. Sequence, Association, & Link Analysis (SAL) Technical Notes. [Online]. Available: http://documentation.statsoft.com/STATISTICAHelp.aspx?path=LinkAnalysis/Overviews/TechnicalNotes
28. Agrawal R. Fast algorithms for mining association rules / R Agrawal, R Srikant // Proceedings of the 20th International Conference on Very Large Data Bases, San Francisco. — 1994. — P. 487—499.

Чертов О.Р., Александрова М.В.

ИСПОЛЬЗОВАНИЕ МЕТОДОВ DATA MINING ДЛЯ АВТОМАТИЧЕСКОГО ФОРМИРОВАНИЯ РЕКОМЕНДАЦИЙ

В статье предложен новый метод автоматического формирования рекомендаций, который использует техники интеллектуального анализа данных в качестве отдельного этапа. Также приведен экспериментальный пример использования предложенного алгоритма.

Ключевые слова: *системы рекомендаций, интеллектуальный анализ данных, поиск ассоциативных правил.*

Chertov O.R., Alexandrova M.V.

DATA MINING METHODS USAGE FOR THE TASK OF AUTOMATED FORMING OF RECOMMENDATIONS

This paper describes a new method for automated forming of recommendations, which uses Data Mining techniques as a separate step. Authors also provide an experimental example of the proposed algorithm usage.

Keywords: *recommender systems, Data Mining, association rules mining.*

Чертов Олег Романович – доцент, кандидат технічних наук, Національний технічний університет України "Київський політехнічний інститут"

Александрова Маргарита Володимирівна – магістрантка, Національний технічний університет України "Київський політехнічний інститут"

Рецензент: Молчанов О. А. **—** д.т.н., професор, завідувач кафедри прикладної математики Національного технічного університету України "Київський політехнічний інститут".