КАТТА МАССИВЛИ МАЪЛУМОТЛАРНИ ИНТЕЛЛЕКТУАЛ ТАҲЛИЛ ҚИЛИШ ЁНДАШУВИ

Бабомурадов Озод Жўраевич
Жиззах шаҳридаги Қозон (Волгабўйи) федерал университети филиали ижрочи директори
bobomuradov@gmail.com



Аннотация

Мазкур мақолада маълумотларнинг интеллектуал таҳлили масалаларига оид кўплаб саволларга жавоб топиш мумкин: Интеллектуал таҳлил қандай амалга оширилади интеллектуал таҳлилдан нима учун ва қандай фойдаланишимиз мумкин? Шубҳасизки, интеллектуал таҳлил усулларини тадбиқ этиш келгусидаги долзарб масалалардан бўлиб қолмоқда. Мақолада ушбу масалалар таҳлил қилинган ҳамда онлайн-дўкондаги транзакциялар асосида мижозлар операцияларини тахлил қилиш алгоритми кўриб чиқилган. Тадқиқотнинг асосий мақсадлардан бири шахсга йўналтирилган оптимал модел ва дастурий войсита ишлаб чиқиш бўлиб, турли ҳил мисоллар орқали кенг ёритилган.

Калит сўзлар: Data Mining (DM), Web Mining, Knowledge Discovery (KD), SAS, IBM, Microsoft, Oracle, Amazon, Alexa, Apple, Siri, Yandex, Alisa.

Аннотация

В этой статье вы сможете найти ответы на многие вопросы, связанные с интеллектуальным анализом данных: Как проводится интеллектуальный анализ? Почему и как мы можем использовать интеллектуальный анализ? Несомненно, применение методов интеллектуального анализа останется одной из актуальных задач в будущем. В статье анализируются эти вопросы и рассматривается алгоритм анализа операций клиентов на основе транзакций в интернет-магазине. Одной из основных задач исследования является разработка человекоориентированной оптимальной модели и программного инструмента, что широко иллюстрируется различными примерами.

Ключевые слова: Data Mining (DM), Web Mining, Knowledge Discovery (KD), SAS, IBM, Microsoft, Oracle, Amazon, Alexa, Apple, Siri, Yandex, Alisa.

Annotation

In this article you will be able to find answers to many questions related to data mining: How is data mining performed? Why and how can we use data mining? Undoubtedly, the use of data mining methods will remain one of the urgent tasks in the future. The article analyzes these

questions and considers an algorithm for analyzing customer transactions based on transactions in an online store. One of the main objectives of the study is to develop a human-oriented optimal model and software tool, which is widely illustrated by various examples.

Keywords: Data Mining (DM), Web Mining, Knowledge Discovery (KD), SAS, IBM, Microsoft, Oracle, Amazon, Alexa, Apple, Siri, Yandex, Alisa.

Кириш. Маълумотларни интеллектуал тахлил қилиш. Республикамизда барча соҳа сингари ахборот коммуникация ва компютер технологияларини ривожлантиришга катта эътибор қаратилмоқда. Ҳозирги кунда вужудга келган обектларни таҳлил қилиш билан боғлиқ ахборотларнинг кун сайин ортиб бориши, янгича усулларнинг кўпайиши маълумотларни интеллектуал таҳлиллаш масалаларининг самарали ечимларини олишга бўлган эҳтиёжни келтириб чиқармоқда[1].

Ҳозирги вақтда жамият ҳаётининг барча жабҳаларида жадал ривожланиш тенденциялари, айниқса, технологик соҳада ҳар қачонгидан ҳам сезиларли. Бу маълумотларни қайта ишлашнинг автоматлаштирилган ва интеллектуал технологияларини, зарур маълумотларни топиш ва ретроспектив маълумотларда яширин нақш ва алоқаларни аниқлаш усулларини ишлаб чиқишни талаб қиладиган ахборот ҳажмининг улкан ўсиши билан боғлиқ. Янги олинган билимлар сифатни бошқариш қарорларини қабул қилиш учун зарур бўлиб, уларнинг аҳамияти ижтимоий муносабатлар ва тизимларнинг тобора мураккаблашиши туфайли ортиб бормоқда. Шу муносабат билан, тизимларнинг ишончлилигини ва уларнинг кўзга кўринмас ўзгаришларга жавоб беришда мослашувчанлигини ошириш ҳозирда жуда муҳимдир. Бироқ, маълумотларнинг бутун мажмуаси орасида одам янги ва фойдали маълумотларнинг фақат кичик бир қисмини ажратиб олишга қодир. Шу сабабли, маълумотларни ишлаб чиқиш ёки DM фаол фойдаланиш олинган фойдали билимлар миқдорини ошириши ва кўпроқ яширин нақшларни аниқлаши мумкин[2].

Реал вақт режимида катта веб-маълумотларнинг интеллектуал таҳлилини ҳал қилишнинг замонавий муаммолари ва истиқболларини, шунингдек, амалий мисол ёрдамида катта веб-маълумотлар учун Web Mining технологиясини амалий татбиқ этиш имкониятларини ўрганиш асосий мақсад қилиб олинди.

Маълумотларни интеллектуал таҳлил қилиш усуллари ва хусусан, DM технологиялари бугунги кунда маълумотлар базасини бошқариш тизимларида инсон фаолиятининг турли соҳаларида қарорлар қабул қилиш учун зарур бўлган маълумотларда илгари номаълум ва амалий жиҳатдан фойдали билимларни топишнинг долзарб муаммосини ҳал қилиш учун кенг қўлланилади.

Асосий қисм. Маълумотларни интеллектуал тахлил қилишнинг асосий таърифлари ва босқичлари. Маълумотни интелллектуал таҳлил қилиш - бу катта маълумотлар тўпламларида фойдали маълумотларни топиш жараёни. Маълумотларни интеллестуал таҳлил қилиш маълумотларда мавжуд нақш(pattern) ва тенденцияларни аниқлаш учун математик таҳлилдан фойдаланади. Одатда, бундай нақш(pattern)ларни аньанавий маълумотларни қидириш орқали аниқлаб бўлмайди, чунки муносабатлар жуда мураккаб ёки маълумотлар ҳажми жуда ката бўлади.

Хом маълумотлар аниқ спецификацияга эга бўлмаган ва қайта ишланмаган (ёки минимал қайта ишланган) маълумотларни ўз ичига олган маълумотлар форматига ишора қилади, бу маълумот йўқолишининг олдини олади[11]. Хорижий адабиётларда маълумотларга дастлабки ишлов бериш атамаси Knowledge KD [12] va DM деб талқин қилинади[13].

Маълумотлар базаларида KD (knowledge discovery) деганда маълумотлардаги ишончли, янги, потенциал фойдали ва яхши талқин қилинадиган нақшларни (тузилмаларни) аниқлашнинг ҳар қандай аҳамиятсиз жараёни тушунилади. Шундай қилиб, KD жараёни маълумотларни тайёрлаш, маълумотлар базасида намуналарни қидириш, олинган билимларни баҳолаш, процедураларни созлаш ва такрорлашни ўз ичига олган кўп босқичли процедуралар тизими сифатида тушунилади[9].

DM маьлумотлар базасидан олинган намуналарни яратиш учун махсус алгоритмларни қўллашдан иборат бўлган KD жараёнининг босқичини англатади. Намуналар тўплами очиқ бўлиши мумкин ва уларни санаб ўтиш махсус алгоритм билан амалга оширилади[10].

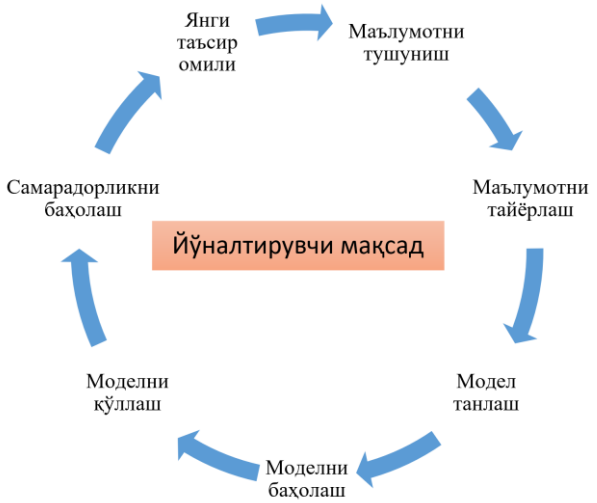
Маълумотлрни интеллектуал таҳлили вазифаларининг ўзига хос хусусияти шундаки, манба маьлумотлари етарли даражада расмийлаштирилмаган, аммо махсус дастурлар ёрдамида ундан янги билимларни олиш мумкин.

Инглиз тилидаги «Data Mining» атамаси ўзбек тилига аниқ таржимага эга эмас (маьлумотлар қазиб олиш, маьлумотлар кашфиёти, ахборотни қазиб олиш, маълумотларни интеллестуал тахлил қилиш) кўп ҳолларда у асл нусхада қўлланилади. 1978 йилда пайдо бўлган ушбу концепция ўзининг замонавий талқинида тахминан 1990-йилларнинг биринчи ярмидан бошлаб юқори машҳурликка эришди. Шу вақтгача маьлумотларни қайта ишлаш ва таҳлил қилиш амалий статистика доирасида амалга оширилди ва кичик маьлумотлар базаларини қайта ишлаш муаммолари асосан ҳал қилинди.

Ушбу атаманинг классик таьрифи 1996 йилда ушбу ёналишнинг асосчиларидан бири Пятецкий-Шапиро томонидан берилган. DM - бу «машина» (алгоритмлар, суньий интеллект воситалари) томонидан илгари номаьлум бўлган, аҳамиятсиз бўлмаган, амалий жиҳатдан фойдали ва инсон талқини учун очиқ бўлган хом маьлумотлардаги яширин билимларни тадқиқ қилиш ва кашф қилишдир [5].

DM маьлумотлар базаларида билимларни кашф қилиш (Knowledge Discovery in Databases, KDD) жараёнининг босқичларидан биридир.

Ташкил этилаётган корпоратив ахборот тизимлари, Миллий ахборот тизимлари, Миллий интернет сегментининг ривожланиши республикадаги ижтимоий-иқтисодий соҳаларда кенг қамровли тарзда катта массивли маьлумотларни таҳлиллашга ё’налтирилган тизимларни жорийлаштириш мақсадга мувофиқ бўлар экан. Бундай воситани қўллашни қуйидаги ёндашувга асосланган ҳолда амалга оширилиши ўринли (2-расм). Ишлов берилаётган маьлумотларнинг бошқарув талабларига мос келувчи механизмини қуришда маьлумотларни бошқариш механизми интеллектуал моделларни ўзида мужассамлаштирган бўлиши керак. Маьлумотлар турли кесимлар учун тайёрлаб олингандан сўнг, унга ишлов бериш учун модел танланади. Моделнинг маьлумотлар билан ишлаши баҳоланиб, уни бирор таҳлил (қарор қабул қилувчи шаҳс талабига кўра) кўриниши амалга оширилади. Шундан сўнг мавжуд натижалар баҳолари асосида самарадорлик аниқланади. Умумий иш сифатини ошириш учун қўшимча таьсир омиллари киритилади. Шу тартибда бизнес жараёнларда интеллектуал таҳлил амалга оширилади[3].



2-расм. Тахлиллаш воситаларини жорийлааштириш ёндошуви.

DM технологиясидан фойдаланишга мисоли тариқасида «Лента» гипермаркет тармоғини олиб кўрамиз. Ушбу усулдан фойдаланган ҳолда мижозларнинг содиқлик карталарини таҳлил қилиб, занжир ассортиментни сезиларли даражада оптималлаштириш ва дисплей ва нархларни бошқаришга муваффақ бўлган. Ҳозирги вақтда маьлумотларни қидириш технологияси истеьмолчиларнинг сўровларини қайта ишлаш ва харидорлар сотиб олишни хоҳлайдиган товарлар ва хизматларни таклиф қилиш учун мўлжалланган тизимларда ҳам фаол фойдаланилмоқда. Ушбу тизимлар маьлумотларни реал вақт режимида таҳлил қиладиган DM усули асосида қурилган. Amazondan Alexa, Appledan Siri ва Yandexдан Alisa овозли ёрдамчилари сўровларнинг 60 фоизини DM усули асосида қайта ишлайди[4].

Онлайн дўкондаги транзакциялар асосида катта веб-маьлумотларни таҳлил қилиш учун бирлашма қоидалари алгоритмини амалий қўллаш мисоли.Веб-маьлумотларни таҳлил қилишнинг амалий муаммоси - Ассоциатця таснифи модели тез-тез содир бўладиган операцияларни аниқлаш учун ишлатилади. Транзакция деганда мижоз томонидан Интернет тармоғида бир сеансда сотиб олинган товарлар тўплами тушунилади.

Берилган: $I = \{i_1, i_2, i_3, \ldots i_n\}$ товарлар тўплами (элементлар) бўлсин. D транзакциялар тўплами бўлсин, $T \subseteq I$ бу йерда ҳар бир транзакция $T\text{-}\{i_1, i_2, i_3, \ldots i_n\}$ элементлар тўплами. Ҳар бир транзакция иккилик вектор ҳисобланади. бу йерда $t[k] = 1$ (агар транзакцияда ик элементи мавжуд бўлса), акс ҳолда $t[k] = 0$. Т транзакцияси Х ни ўз ичига олади, агар $X \subset T$ бўлса, I дан баьзи элементлар тўплами олинади.

Ассоциация қоидаси импликация сифатида аниқланади, $X \Rightarrow Y$ импликация, бу эрда $X \subset I$, $Y \subset I$ ва $X \cap Y = \emptyset$. $X \Rightarrow Y$ қоидасининг қўллаб-қувватланиши support c миқдоридир, агар транзакция-ларнинг c% D дан бўлса, таркибида $X \cup Y$ мавжуд.

$$\mathrm{conf}(X \Rightarrow Y) = \sup p(X \cup Y)$$

Қоиданинг ҳақиқийлиги Х нинг Y ни назарда тутиши эҳтимолини белгилайди, $X \Rightarrow Y$ қоидасининг ишончлилиги confidence c қийматига тенг, агар транзакцияларнинг c% D дан бўлса, Х ни ҳам Y ни ҳам ўз ичига олади формула (1) бўйича:

$$\mathrm{conf}(X \Rightarrow Y) = \frac{\sup p(X \cup Y)}{\sup p(X)} \qquad (1)$$

Алгоритмнинг хусусиятлари ишончлилик ва қоидаларни қўллаб-қувватлашдир.

Амалий мисол. Буюртмаларнинг 70 фоизи чой ва асал арралашмасини сотиб олган. Барча буюртмаларнинг 8% ҳам чой, ҳам асал аралашмасини ўз ичига олади.

8% - қўллаб-қувватлаш;

70% - "Чой" → "Асал сметаси" тахминан 70% эҳтимоллик билан. Бу қоиданинг ҳақиқийлиги.

Хулоса: алгоритмнинг мақсади ўзаро муносабатларни аниқлашдир: агар транзакцияда Х компонентлар тўплами учраса, у ҳолда Y компонентлар тўплами худди шу транзакцияда пайдо бўлиши керак деган хулосага келишимиз мумкин[6].

Ассоциация қоидаси алгоритмлари қоиданинг қо‘ллаб-қувватланиши ва ҳақиқийлиги [7] бо‘йича чекловларга эга. Кўпинча қоидалар сони белгиланган чегаралар (threshold) билан чекланиши керак: минимал қўллаб-қувватлаш ва ишончлилик - minsupp ва minconf.

Агар қўллаб-қувватлаш қиймати юқори бўлса, таҳлил пайтида аниқ қоидалар топилади. Агар қиймат паст бўлса, таҳлил аниқ бўлмаган ва асоссиз бўлган жуда кўп қоидаларни ўз ичига олади.

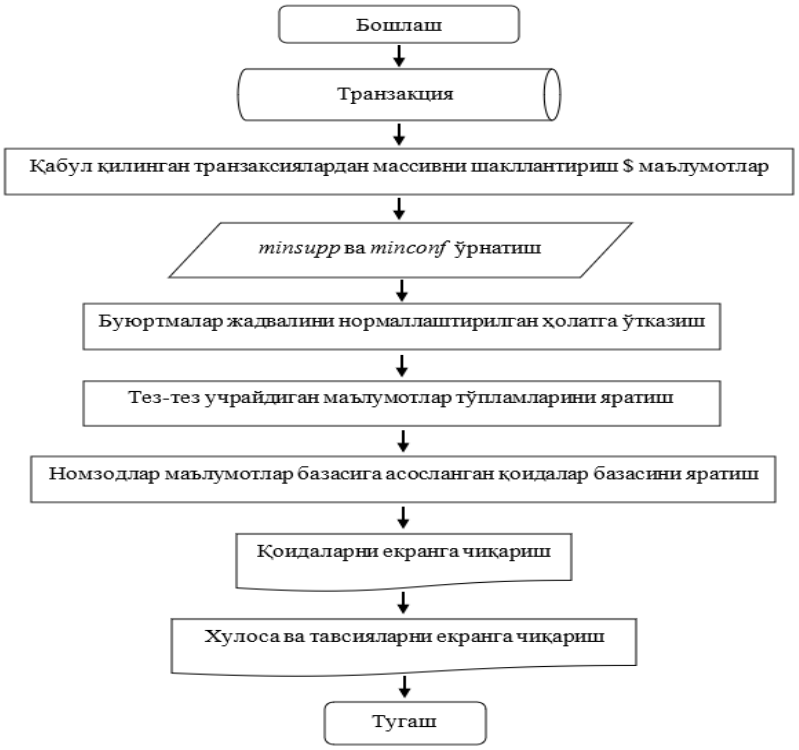
Алгоритм икки босқичда ишлайди:

- излаш ва minsupp threshold (минимал қўллаб-қувватлаш чегараси) қониқтирувчи элементлар мажмуини шакллантириш. Ушбу элементлар тўплами тез-тез учрайдиган дейилади
- minconf threshold га мос келадиган аниқлик билан ўрнатилган компонентлар тўпламидан қоидалар яратиш.

Ассоциация қоидалари алгоритмининг айрим турлари қўшимча миқдорларни киритади. Бу миқдорлардан бири муҳимлик (importance) [8]. Ушбу қиймат (2) формула бўйича ҳисобланади:

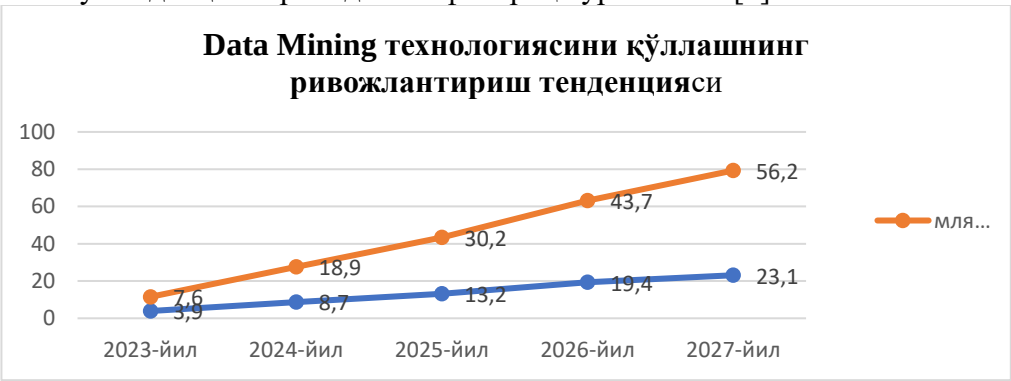
$$\mathrm{Imp}(X \Rightarrow Y) = \log\left(\frac{p(X|Y)}{p(X|notY)}\right) \qquad (2)$$

Амалда ишонч параметридан кўра муҳимлик параметри кўпроқ қўлланилади. Муҳим параметр Y маҳсулоти X маҳсулоти харид қилинганда пайдо бўлиш эҳтимолидаги ўзгаришларни кузатишга ёрдам беради.



1-расм. Алгоритмнинг блок схемаси

Тақдим этилган мисоллар ва йирик корпорациялар: SAS, IBM, Microsoft, Oracle ва бошқаларнинг фаолиятидан DM бозорининг доимий ўсишини кўришимиз мумкин. Мутахассисларнинг фикрига кўра, 2027 йилга бориб маълумотларни қидириш технологияларидан фойдаланиш тенденцияси 23,1 фоизга ўсиб, 56,2 миллиард долларга йетади. Ушбу тенденция 3-расмда батафсилроқ кўрсатилган[4].



3-расм. Маълумотларни интеллектуал тахлиллаш технологиясининг ривожланиш тенденцияси[4].

Агар X сотиб олинган бўлса, у ҳолда Y маҳсулоти, агар муҳимлик параметри ижобий бўлса, сотиб олишнинг энг юқори эҳтимоли бор. Агар муҳимлик параметри салбий бўлса, у ҳолда Y X билан бирга сотиб олинадиган энг кам эҳтимолга эга. Агар муҳимлик параметри нолга тенг бўлса, у ҳолда маҳсулотлар бир-бирига боғлиқ эмас.

Мижозлар операцияларини таҳлил қилиш алгоритмини қуйидаги блок схемада кўриб ўтамиз 1-расм.

Хулоса.

DM технологияси ҳали ҳам амалий нуқтаи назардан анча мураккаб ва назарий жиҳатдан қўшимча ўрганишни талаб қилади. Ушбу алгоритмни сайтларни бошқариш тизимига татбиқ этиш бўйича ўтказилган ҳисоблаш эксперименти натижалари шуни кўрсатдики, ишлаб чиқилган модулнинг динамик веб-илова билан интеграциялашуви соддалаштирилган, онлайн-дўкондаги транзакциялар асосида веб-маълумотларни таҳлил қилиш харажатлари ва вақти қисқарган.

Маьлумотлардаги нақшларни аниқлашнинг асосий муаммоси ахборот массивларини саралаш учун кетадиган вақтдир. Маьлум усуллар бундай қидирувни суньий равишда чеклайди ёки қидирув самарадорлигини камайтирадиган бутун қарор дарахтларини яратади. Ушбу муаммони ҳал қилиш маьлумотларни ишлаб чиқиш маҳсулотларини ишлаб чиқувчиларнинг асосий мақсади бўлиб қолмоқда.

Фойдаланилган адабиётлар.

1. Mirzaakbar Xudayberdiyev, Oybek Qoraboshev, Nurmuhammad Alimqulov, Sherzod Yuldashev MA’LUMOTLARNI INTELLEKTUAL TAHLIL QILISH ASOSIDA FAVQULODDA VAZIYATLARGA OID AXBOROTLARGA ISHLOV BERISH // AI. 2022. №2. URL: https://cyberleninka.ru/article/n/ma-lumotlarni-intellektual-tahlil-qilish-asosida-favqulodda-vaziyatlarga-oid-axborotlarga-ishlov-berish.
2. Бохан Петр Андреевич ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ // StudNet. 2022. №6. URL: https://cyberleninka.ru/article/n/intellektualnyy-analiz-dannyh-1.
3. Озод Жўраевич Бабомурадов, and Шохрух Маткарим ўғли Матчонов. “BI-ТИЗИМЛАР ВА УЛАРНИ АМАЛИЙ МАСАЛАЛАРНИ ЕЧИШДА ҚЎЛЛАШ”. RESEARCH AND EDUCATION 1, no. 1 (April 30, 2022): 212–234. https://researchedu.org/index.php/re/article/view/706.
4. РБК * Тренды // Банки, ретейл, медицина: кто использует Data Mining и для чего [Электронный ресурс]. Режим доступа — https://trends.rbc.ru/trends/industry/61b359739a7947c7376ef7ce (Дата обращения: 25.03.2022).
5. Мулюкова К. В., Курейчик В. М. Построение ассоциативной классификационной модели данных на основе метода Apriori // Открытое образование. 2020. №4. URL: https://cyberleninka.ru/article/n/postroenie-assotsiativnoy-klassifikatsionnoy-modeli-dannyh-na-osnove-metoda-apriori.
6. Пивоварова Наталья Владимировна, Видунова Светлана Игоревна Интеллектуальный анализ данных в фармацевтическом бизнесе // Вестник евразийской науки. 2016. №6 (37). URL: https://cyberleninka.ru/article/n/intellektualnyy-analiz-dannyh-v-farmatsevticheskom-biznese.
7. В.А. Биллиг, О.В. Иванова, Н.А. Царегородцев Построение ассоциативных правил в задаче медицинской диагностики // Программные продукты и системы. 2016. №2 (114). URL: https://cyberleninka.ru/article/n/postroenie-assotsiativnyh-pravil-v-zadache-meditsinskoy-diagnostiki.
8. Олянич И.А. Сравнение алгоритмов построения ассоциативных правил на основе набора данных покупательских транзакций // Известия Самарского научного центра Российской академии наук. 2018. № 6-2 (20). С. 379–382. URL: http://www.ssc.smr.ru/media/journals/izvestia/2018/2018_6_379_382.pdf
9. Финн В.К. Об интеллектуальном анализе данных // Новости искусственного интеллекта, №3, 2004, С. 3-19.

10. Piatetsky-Shapiro G, Frawley W J. Knowledge Discovery in Databases. USA: MIT Press, 1991.

11. Agrawal R., Mannila H., Srikant R., Toivonen H. and Verkamo I. Fast Discovery of Association Rules. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Menlo Park, Calif.: AAAI Press, 1996, pp. 307-328.