

В роботі розглянута теоретико-множинна модель семантичних полів в масивах текстових даних. Проаналізовано утворення семантичних полів на основі лексемних відношень. Показана ефективність використання концепції семантичних полів у векторній моделі текстових даних

Ключові слова: інтелектуальний аналіз текстів, модель векторного простору

В работе рассмотрена теоретико-множественная модель семантических полей в массивах текстовых данных. Проанализированы образования семантических полей на основе лексемных отношений. Показана эффективность использования концепции семантических полей в векторной модели текстовых данных

Ключевые слова: интеллектуальный анализ текстов, модель векторного пространства

The theoretical set model of semantic fields in text arrays was considered in this work. The forming of semantic fields on the base of lexeme relations was analyzed. The effectiveness of the use of semantic fields in the vector model of text data was shown

Keywords: text mining, vector space model

ВИКОРИСТАННЯ КОНЦЕПЦІЇ СЕМАНТИЧНОГО ПОЛЯ У ВЕКТОРНІЙ МОДЕЛІ ТЕКСТОВИХ ДОКУМЕНТІВ

Б.М. Павлишенко

Кандидат фізико-математичних наук, доцент
Факультет електроніки
Львівський національний університет
імені Івана Франка
вул. Драгоманова, 50, м. Львів, 79005
E-mail: pavlish@yahoo.com

Вступ

В сучасних алгоритмах аналізу текстів часто використовують модель векторного простору. Основна ідея цієї моделі полягає у представленні кожного текстового документа у вигляді вектора у деякому векторному просторі [1,2]. Вважають, що точки, які є близькі між собою в цьому просторі відображають семантично близькі документи і навпаки - семантично близькі документи відображаються близькими точками у фазовому просторі. Деякий запит на пошук документів також можна розглядати як точку у векторному просторі документів. Результат такого запиту можна представити як список документів відсортованих в порядку зростання відстані від точки запиту до точок, які представляють документи.

Одним із застосувань векторної моделі є автоматизоване отримання формалізованих знань. Більшість пошукових машин використовують векторну модель у своїх алгоритмах для визначення подібності між запитом та документами. В основі використання векторної моделі лежить статистична гіпотеза, яка полягає в тому, що статистичні характеристики використання слів відображають ті поняття, які люди мають на увазі в цих текстах [2]. Одним із методів представлення масиву документів у векторному просторі є організація векторів документів у вигляді матриці текстових частот типу лексеми-документи. Рядки таких матриць відповідають лексемам, а стовпці є векторами відповідних документів. В аналізі текстів розглядають гіпотезу мультимножини лексем, яка полягає в тому, що можна розрахувати відповідність документів до запиту представляючи документ та запит як мультимножи-

ну лексем. Сукупність частот рядка частотної матриці представляє образ деякої лексеми і характеризує її використання в масиві текстів. Сукупність частот стовпця визначає образ деякого документа. Подібність між документами та запитом можна аналізувати представляючи запит як псевдо-документ та розглядаючи вектори рядків матриці лексеми-документи [2].

Аналізуючи подібність між векторами-рядками матриці можна визначити подібність між лексемами документів. В аналізі подібності лексем замість текстових документів також розглядають їх складові - фрази, речення, параграфи, розділи.

Розглянемо ряд гіпотез, які лежать в основі векторної моделі текстів [2]. Статистична семантична гіпотеза припускає, що статистичні характеристики вживання лексем можуть бути використані для визначення змісту сказаного. Якщо деякі частини тексту мають подібні вектори в частотних матрицях тоді ці частини мають подібні значення. Гіпотеза сукупності лексем говорить про те, що частоти лексем в документі відображають зв'язок між документом та запитом. Якщо документ та псевдодокумент мають подібні стовпці в частотній матриці, тоді вони мають подібні значення.

Дистрибутивна гіпотеза припускає, що лексеми, які зустрічаються в подібних контекстах мають подібні значення. Якщо лексеми мають подібні вектори-рядки в частотній матриці, тоді вони мають подібні значення. Розширена дистрибутивна гіпотеза говорить, що лексеми, в яких зустрічаються подібні пари лексем мають подібні значення. Якщо контексти мають подібні вектори-рядки в частотній матриці лексемна_пара-контекст значить вони відображають подібні значення. Гіпотеза латентних відношень при-

пускає, що пари лексем, які зустрічаються в подібних лексемних шаблонах, мають подібні семантичні зв'язки. Якщо лексемні пари мають подібні вектори-рядки в частотній матриці лексемна_пара-шаблон, тоді вони мають подібні семантичні зв'язки.

Однак, векторна модель текстових документів має ряд проблемних місць. Зокрема, аналіз всіх лексем текстового словника приводить до значного обсягу обчислень внаслідок великого розміру матриці лексеми-документи. Крім того, текстові документи та псевдо-документи можуть бути близькими у просторі, базис якого утворений лексемними групами. Такими групами можуть бути, зокрема, семантичні поля. Під семантичним полем розуміють таку множину лексем, які об'єднані певним спільним поняттям [3,4]. Прикладом семантичних полів може бути поле руху, поле комунікації, поле сприйняття та інші.

Постановка задачі

Розглянемо теоретико-множинну концепцію семантичного поля. Проаналізуємо утворення семантичних полів як сегментів семантичного розбиття лексемного словника. Розглянемо формування лексемного складу на основі лексемних відношень в текстових масивах. Проаналізуємо використання теоретико-множинної моделі семантичних полів у векторному представленні текстових документів.

Теоретико-множинна концепція семантичного поля

Розглянемо утворення поняття «семантичне поле» в процесі аналізу текстових даних. Нехай існує деякий словник лексем, які зустрічаються в аналізованих текстових масивах. Опишемо цей словник як впорядковану множину

$$W = \{ w_i \mid i = 1, 2, \dots, N_w \}, \tag{1}$$

де N_w - кількість лексем в словнику. Введемо множину семантичних полів

$$S = \{ S_k \mid k = 1, 2, \dots, N_s \}, \tag{2}$$

де N_s - кількість семантичних полів. Семантичні ознаки лексем будемо характеризувати відображенням

$$U_{WS}: W \rightarrow S, w_i \rightarrow s_k, i = 1, 2, \dots, N_w; k = 1, 2, \dots, N_s \tag{3}$$

Тобто у відповідність кожній лексемі ставлять деякий елемент множини S . Множина значень S може мати різну природу, наприклад це може бути множина назв деяких семантичних класів. Шкала семантичних ознак є номінальною, якщо лексеми набувають деяких назв із множини S . Номінальна шкала володіє класифікаційним потенціалом, коли за допомогою відображення (3) можна утворити групування елементів множини W , які мають спільні назви із множини S . В загальному, класифікацію лексем за семантичними полями будемо розглядати

як відображення множини лексем на множину семантичних полів. Семантичну класифікацію розглянемо як деяку сукупність відображень лексем на множину дійсних чисел. Можливу квантифікацію лексемних відображень можна пов'язати із частотами лексем в текстових об'єктах. Розглянемо утворення семантичного поля на основі відношення еквівалентності. Нехай існує деяке бінарне відношення

$$S_k^b \subseteq W \times W \tag{4}$$

Розглянемо деяку квантитативну ознаку лексеми $x_k^s(w_i)$, яка кількісно характеризує лексемні відношення заданого типу у множині аналізованих текстових об'єктів. Наприклад, це може бути частота появи лексеми w_i в заданому лексемному шаблоні.

Пов'яжемо із ознакою $x_k^s(w_i)$ бінарне відношення

$$S_k^b = \{ (w_i, w_j) \mid x_k^s(w_i) = x_k^s(w_j) \} \tag{5}$$

Можна показати, що відношення S_k^b є рефлексивним, тобто

$$(w_i, w_i) \in S_k^b, \forall w_i \in W, \tag{6}$$

симетричним, тобто

$$(w_i, w_j) \in S_k^b \Rightarrow (w_j, w_i) \in S_k^b, \forall w_i, w_j \in W \tag{7}$$

і транзитивним, тобто

$$(w_i, w_j) \in S_k^b, (w_j, w_l) \in S_k^b \Rightarrow (w_i, w_l) \in S_k^b, \forall w_i, w_j, w_l \in W \tag{8}$$

Рефлексивне, симетричне і транзитивне відношення називають еквівалентністю [5]. Еквівалентність S_k^b повністю характеризує породжуючи його ознаку $x_k^s(w_i)$ і дає можливість визначити множину лексем, які не розрізняють за цією ознакою:

$$S_k^c = \{ w_i \mid (w_i, w_j) \in S_k^b \} \tag{9}$$

Якщо S_k^c є деяким семантичним відношенням, тоді неспівпадаючі множини S_k^c утворюють розбиття лексемного словника W на семантичні класи

$$S_{sc} = \{ S_k^c \mid k = 1, 2, \dots, N_s \}. \tag{10}$$

Бінарне відношення S_k^b може також породжуватись деяким логічним висловлюванням $Q(w_i, w_j)$

$$S_k^b = \{ (w_i, w_j) \mid Q(w_i, w_j) = \text{true} \} \tag{11}$$

де $Q(w_i, w_j)$ описує деяку умову, наприклад одночасне використання в текстових шаблонах заданої структури. Умова породження бінарного відношення S_k^b може також описуватись деяким правилом підстановки в заданій схемі формальної граматики. Таке правило може бути сформована деяким регулярним виразом.

Розглянемо рангову ознаку $x_k^{rs}(w_i)$, яка утворює бінарне відношення

$$S_k^{sb} = \{ (w_i, w_j) | x_k^s(w_i) \leq x_k^s(w_j) \} \quad (12)$$

Можна показати, що таке бінарне відношення є рефлексивне, транзитивне та лінійне. Такі відношення називають лінійними квазіпорядками [5]. Квазіпорядок S_k^{sb} породжує рангову шкалу семантичного поля S_k^r . У випадку формування семантичного поля за допомогою рангових ознак можна визначити внутрішню структуру поля, для якої можна сформувати внутрішній частковий порядок виділивши структурні групи всередині семантичного поля. Такими групами можуть бути, наприклад, частотне ядро семантичного поля, основна частотна область, периферійна частотна область. Для кожної із цих груп можна визначити умови для семантичної ознаки, за якою лексеми всередині цих груп не розрізняються.

Відношення еквівалентності та квазіпорядку визначають номінальні та рангові семантичні шкали для лексемного складу словника текстових масивів на основі лексемних відношень елементів різних класів семантичного розбиття.

В семантичному розбитті словника можна виявити відповідні структурні зв'язки. Семантичним розбиттям із структурою назвемо пару (S^W, Z_{str}) , де S^W - є розбиття лексемного словника на відповідні семантичні класи. Відношення

$$Z_{str} \subseteq \{1, 2, \dots, N_{sc}\} \times \{1, 2, \dots, N_{sc}\}, \quad (13)$$

є відношенням відповідних зв'язків між семантичними класами, де N_{sc} - кількість семантичних класів розбиття. Умова $(i, j) \in Z_{str}$ означає, що існує зв'язок між лексемами семантичних класів S_i^W та S_j^W . Відношення Z_{str} можна представити за допомогою булевої матриці $M^{str} = \{m_{ij}^{str}\}$ вважаючи, що $m_{ij}^{str} = 1$ тоді, коли $(i, j) \in Z_{str}$. Аналізуючи матрицю M^{str} можна виявити сегменти розбиття, тобто групи класів, які пов'язані між собою за допомогою розбиття Z_{str} . Під сегментом розбиття розуміють деяку множину утворену об'єднанням класів розбиття. Здійснюючи перестановку рядків та стовпців цієї матриці можна виявити підматриці, які заповнені одиницями. Індеси стовпців та рядків будуть описувати індеси семантичних класів, які можна віднести до єдиного спільного сегменту. Такі підматриці формують бінарні кластери, тобто групи рядків та стовпців, об'єднаних спільними властивостями. В цьому випадку елементи матриці, які відповідають цим рядкам та стовпцям рівні одиниці. Визначимо бінарний кластер в структурному відношенні Z_{str} . Нехай X_{str}^Z є множина рядків матриці Z_{str} , а Y_{str}^Z - множина стовпців. Матрицю Z_{str} позначимо так

$$Z_{str} = (X_{str}^Z, Y_{str}^Z). \quad (14)$$

Підмножини

$$I_{str}^Z \subseteq X_{str}^Z, J_{str}^Z \subseteq Y_{str}^Z \quad (15)$$

утворюють підматрицю $M_{bc} = (I_{str}^Z, J_{str}^Z)$, яка описує бінарний кластер за рядками та стовпцями підматриці, якщо виконується умова

$$M_{bc} = \{ m_{ij}^{str} | m_{ij}^{str} = 1, i \in I_{str}^Z, j \in J_{str}^Z \} \quad (16)$$

Враховуючи (14)-(16) семантичний сегмент можна визначити, наприклад, так

$$S_k = \{ S_i^R | i \in (I_{str}^Z \cup J_{str}^Z) \} \quad (17)$$

Виходячи із наведеного аналізу семантичним полем назвемо сегмент, який утворюється семантичними класами об'єднаними бінарним кластером в структурному відношенні семантичного розбиття лексемного словника текстових масивів. Отже, за допомогою пари (S^W, Z_{str}) , яка описує розбиття словника на семантичні класи із структурою, можна задавати семантичні поля лексемного словника.

Векторна модель текстових документів

Основним методом формування множини лексем семантичного поля є експертний лексикографічний аналіз. Розглянемо можливість формування семантичного поля на основі лексемних відношень в текстових масивах. Сукупність текстових документів опишемо такою множиною

$$D = \{ d_j | j = 0, 1, 2, \dots, N_d \}, \quad (18)$$

де N_d - кількість документів. Під документом з $j=0$, будемо вважати документ з нейтральним текстом, який відповідає лінгвостатистичні нормі. Документ d_j з множини текстових документів D можна представити як упорядковану множину слів, порядок елементів якої відповідає порядку слів у цьому документі

$$T_j^d = \{ t_{lj} | l = 1, 2, \dots, N_l^j \} \quad (19)$$

Упорядкований за алфавітом словник текстового документа d_j розглянемо як мультимножину W_j^d над множиною словника W

$$W_j^d = \{ n_{ij}^{wd} (w_i) | w_i \in d_j, i = 1, 2, \dots, N_w \} \quad (20)$$

де n_{ij}^{wd} - кількість входжень лексеми w_i із словника W в множину лексем текстового документа d_j , яку можна визначити як

$$n_{ij}^{wd} = \sum_{l=1}^{N_l^j} f_{wd}(t_{lj}, w_i), \quad (21)$$

$$\text{де } f_{wd}(t_{lj}, w_i) = \begin{cases} 1, & t_{lj} = w_i \\ 0, & w_{lj}^d \neq w_i \end{cases} \quad (22)$$

Відображення лексемного складу словника W на множину семантичних полів S (3) задамо таблицею, яка визначається експертним лексикографічним аналізом [3,4]. Лексемний склад семантичного поля s_k визначимо як

$$W_k^s = \left\{ w_i | w_i \xrightarrow{U_{ws}} s_k, i = 1, 2, \dots, N_w \right\} \quad (23)$$

Множину образів відображення U_{ws} розглянемо як мультимножину над множиною семантичних полів S

$$S_i = \{ n_k^s(s_k) | k = 1, 2, \dots, N_s \} \quad (24)$$

де n_k^s - кількість лексем словника W , які відносять до семантичного поля s_k :

$$n_k^s = \sum_{i=1}^{N_w} f_s(w_i, s_k), \quad (25)$$

$$\text{де } f_s(w_i, s_k) = \begin{cases} 1, & w_i \in W_k^s \\ 0, & w_i \notin W_k^s \end{cases} \quad (26)$$

Введемо мультимножину образів відображення U_{ws} семантичних полів для окремого документа d_j

$$S_j^d = \{ n_{kj}^{sd}(s_k) | k = 1, 2, \dots, N_s \} \quad (27)$$

де n_{kj}^{sd} - кількість лексем семантичного поля s_k в лексемному складі документа d_j

$$n_{kj}^{sd} = \sum_{i=1}^{N_j^t} f_s(t_{ij}, s_k), \quad (28)$$

$$\text{де } f_s(t_{ij}, s_k) = \begin{cases} 1, & t_{ij} \in W_k^s \\ 0, & t_{ij} \notin W_k^s \end{cases} \quad (29)$$

Введемо оператор відображення лексемного словника W на множину квантитативних ознак в масиві документів

$$U_{wd} : w_i \rightarrow p_{ij}^{wd}, \quad i = 1, 2, \dots, N_w, j = 1, 2, \dots, N_d \quad (30)$$

У загальному випадку величина p_{ij}^{wd} може мати довільне походження квантитативної характеристики. У подальшому будемо розглядати цю величину як текстову частоту лексеми w_i у текстовому документі d_j , яка визначається такою функціональною залежністю

$$p_{ij}^{wd} = \frac{n_{ij}^{wd}}{N_j^t}. \quad (31)$$

Аналогічно введемо оператор відображення семантичного складу S_j^d текстового документа d_j на множину квантитативних ознак:

$$U_{sd} : s_k \rightarrow p_{kj}^{sd}, \quad k = 1, 2, \dots, N_s, j = 1, 2, \dots, N_d \quad (32)$$

Величина p_{kj}^{sd} визначає структурну частоту лексем семантичного поля s_k у текстовому документі d_j . Визначимо p_{kj}^{sd} за такою формулою

$$p_{kj}^{sd} = \sum_{i=1}^{N_w} p_{ij}^{wd} f_s(w_i, s_k), \quad (33)$$

$$\text{де } f_s(w_i, s_k) = \begin{cases} 1, & w_i \in W_k^s \\ 0, & w_i \notin W_k^s \end{cases} \quad (34)$$

Сукупність значень p_{ij}^{wd} утворює матрицю типу ознака-документ

$$M_{wd} = (p_{ij}^{wd})_{i=1, j=1}^{N_w, N_d} \quad (35)$$

В матриці M_{wd} роль ознаки відіграє текстова частота лексеми. Введемо вектор

$$V_j^w = (p_{1j}^{wd}, p_{2j}^{wd}, \dots, p_{N_w, j}^{wd}) \quad (36)$$

Такий вектор відображає документ d_j в N_w -мірному просторі текстових документів. Сукупність значень p_{kj}^{sd} утворюють іншу матрицю ознака-документ, у якій ознаками виступають частоти семантичних полів у документах:

$$M_{sd} = (p_{kj}^{sd})_{k=1, j=1}^{N_s, N_d}. \quad (37)$$

Вектор

$$V_j^s = (p_{1j}^{sd}, p_{2j}^{sd}, \dots, p_{N_s, j}^{sd}) \quad (38)$$

відображає документ d_j в N_s -мірному просторі текстових документів. Використання векторного представлення (38) дає можливість пошуку подібних документів та псевдодокументів у векторному просторі із базисом, утвореним частотними характеристиками семантичних полів. Цей базис має суттєво меншу розмірність у порівнянні із базисом, утвореним частотними характеристиками лексем словника текстових масивів. Це дає можливість зменшити кількість необхідних обчислень в алгоритмах аналізу текстів.

Утворення семантичних полів на основі лексемних відношень

Розглянемо можливість утворення семантичних полів на основі дистрибутивної гіпотези та гіпотези латентних відношень. Визначимо множину W_k^{sc} деяких базових лексем, які будуть виконувати роль ядра, яке утворює семантичне поле на основі деякого перетворення. Вміст семантичного поля буде утворюватись лексемами, які мають задане відношення до лексем множини W_k^{sc} . Такі відношення можуть бути, наприклад, зумовлені утворенням лексемних пар в тексті лексем семантичного поля із лексемами заданого ядра, або приналежністю лексеми семантичного поля і лексеми полеутворюючого ядра до спільних логічних сегментів текстів. Такими логічними сегментами можуть бути, наприклад, речення, абзаци, розділи. Розглянемо спочатку утворення семантичних полів на основі лексемних пар. Розглянемо текст як ланцюжок лексем

$$T = l_1 l_2 \dots l_{N_t} \quad (39)$$

Вважаємо, що множина лексем в тексті є впорядкована за номером лексеми в тексті. Розглянемо деяке відношення

$$I_k^s \subseteq W \times W \quad (40)$$

Пара (w_i, w_j) належить відношенню I_k^s , якщо першалеξεма w_i належить до полеутворюючого ядра W_k^{sc} , а друга лексема w_j утворює із нею пару лексемного сполучення в тексті, тобто виконується умова

$$(w_i, w_j) \in I_k^s, \quad \text{якщо } w_i \in T, w_j \in T, w_i \in W_k^{sc}, w_i < w_j. \quad (41)$$

Необхідно врахувати, що у великих текстових масивах можуть зустрічатись деякі лексемні пари, які не відображають стійких семантичних зв'язків. Тому необхідно визначити додаткову умову, за допомогою якої можна було б вилучати такі пари із розгляду. Аналогічно до теорії частих множин [6,7], де розглядається поняття підтримки частих множин, розглянемо поняття підтримки лексемних сполучень. Підтримкою лексемного сполучення для лексеми w_j будемо вважати відношення кількості лексемних сполучень цієї лексеми із лексемами полеутворюючого ядра до загальної кількості появ цієї лексеми в аналізованому текстовому масиві

$$\text{Supp}(w_j) = \frac{n((w_i, w_j) \in I_k^s)}{n(w_j)} \quad (42)$$

До подальшого розгляду будемо брати лексеми для яких величина підтримки $\text{Supp}(w_j)$ буде перевищувати деяке наперед задане число

$$\text{Supp}(w_j) \geq \text{Supp}_{\min} \quad (43)$$

Враховуючи (41)-(43), семантичне поле можна визначити як

$$S_k = \{ w_j \mid \exists w_i \in W_k^{sc}, (w_i, w_j) \in I_k^s, \text{Supp}(w_j) \geq \text{Supp}_{\min} \} \quad (44)$$

Аналогічно можна визначити семантичне поле на основі умови спільної появи лексеми із множини полеутворюючого ядра та лексеми семантичного поля в деякому заданому логічному сегменті тексту, наприклад реченні. Розглянемо текст як впорядковану множину логічних сегментів

$$T = \{ g_1, g_2, \dots, g_{N_g} \} \quad (45)$$

В цьому випадку семантичне поле можна розглядати як таку множину

$$S_k = \{ w_j \mid \exists w_i \in W_k^{sc}, \exists g_m \in T, (w_i, w_j) \in g_m, \text{Supp}_{sg}(w_j) \geq \text{Supp}_{\min} \} \quad (46)$$

Підтримка $\text{Supp}_{sg}(w_j)$ має аналогічний зміст до підтримки лексемних сполучень (43), замість кількості лексемних сполучень в тексті розглядається кількість логічних текстових сегментів. Розглянуті лексемні відношення можуть бути узагальнені для випадку лексемних відношень у складному текстовому шаблоні із заданою множиною додаткових умов.

Висновки

Розглянута теоретико-множинна концепція семантичних полів в масивах текстових даних. Семантичні класи утворюються як відношення еквівалентності. Семантичне поле визначається як сегмент, який утворюється семантичними класами, об'єднаними бінарним кластером в структурному відношенні семантичного розбиття лексемного словника текстових масивів. Розглянуто відношення, яке описує розбиття словника на семантичні класи із структурою, яка визначає семантичні поля лексемного словника. Проаналізовано утворення семантичних полів на основі лексемних відношень, зокрема, таких як сполучення в тексті лексем семантичного поля та лексем полеутворюючої множини. Використання концепції семантичних полів є ефективним у векторній моделі текстових документів внаслідок зменшення розмірності фазового простору представлення документів.

Це дає можливість зменшити кількість необхідних обчислень в алгоритмах аналізу текстових даних.

В подальших дослідженнях розглянемо використання моделі семантичних полів у векторному представленні текстових документів в алгоритмах інтелектуального аналізу, зокрема, в алгоритмах кластеризації, пошуку асоціативних зв'язків, та аналізі частих лексемних множин. Цікавим, на нашу думку, є також розгляд моделі семантичних полів у рамках теорії аналізу формальних понять в текстових масивах.

Література

1. Брасегян А.А. Анализ данных и процессов: учеб. Пособие / А.А.Брасегян, М.С.Куприянов, И.И.Холод, М.Д.Тесс, С.И.Елизаров. – СПб.:БХВ–Петербург, 2009. – 512с.:ил.
2. Pantel P. From Frequency to Meaning: Vector Space Models of Semantics / Patrick Pantel, Peter D. Turney // Journal of Artificial Intelligence Research. – 2010. – vol.37. – pp.141-188.
3. Левицкий В.В. Экспериментальные методы в семасиологии / В.В. Левицкий, И.А. Стернин. – Воронеж: Изд-во ВГУ, 1989. – 192с.
4. Вердиева З.Н. Семантические поля в современном английском языке. / З.Н. Вердиева – М.: Высшая школа, 1986. – 120с.
5. Миркин Б.Г. Анализ качественных признаков и структур. / Б.Г. Миркин. – М.: Статистика, 1980. – 319с., ил.
6. Agrawal R. Fast algorithm for mining association rules / R. Agrawal, R. Srikant // Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94), pages 487-499, Santiago, Chile, Sept. 1994.
7. Mannila H. Efficient algorithms for discovering association rules / H.Mannila, H.Toivonen, and A.J.Verikamo // Proc. AAAI'94 Workshop Knowledge Discovery in Databases (KDD'94), pages 181-192, Seattle, WA, July, 1994.