УДК 004.622

ПИТАННЯ ОЧИСТКИ ДАНИХ ПРИ СТВОРЕННІ АВТОМАТИЗОВАНИХ СИСТЕМ НОРМАТИВНО-ДОВІДКОВОЇ ІНФОРМАЦІЇ

© Гула О.Ю.

ТОВ «Ер-Джі-Дейта»
Україна 03056, Київ, вул. Політехнічна, 33, к. 616
Тел.: +380 44 241 9131; факс: +380 44 236 3188

e-mail: *alexg@rgdata.com.ua*

The article describes methods of data identification in referenced data systems/ Algorithm of hierarchical classifiers data consolidation is suggested. Structure of data processing application for unified hierarchical classifier construction is proposed.

Вступ

Нормативно-довідкова інформація (НДІ) в автоматизованих системах призначена для групування та систематизації інформації про поняття, об'єкти, явища, тощо в стандартний формат, що допомагає визначити їх подібність. Систематизація інформації здійснюється шляхом її класифікації, а саме поділом множини об'єктів на підмножини, зі застосуванням прийнятих методів визначення подібності на підставі їх схожості чи несхожості [1]. При ієрархічній системі множина об'єктів ділиться на великі групи, потім кожна група ділиться на ряд множину підгруп, які в свою чергу також можуть ділитися, поступово конкретизуючи об'єкт.

В даній роботі розглядається задача побудови зведеного ієрархічного реєстру товарно-матеріальних цінностей (ТМЦ) корпорації, що об'єднує декілька окремих класифікаторів ТМЦ філій корпорації в єдину централізовану систему НДІ для подальшої обробки експертами.

При створенні такої централізованої системи НДІ значне місце займають питання консолідації, очистки даних та їх узгодження [2, 3], зокрема:

1. Усунення дублювання як окремих об'єктів, так і цілих груп об'єктів.
2. Усунення помилок та розбіжностей в значеннях атрибутів об'єктів (наприклад, помилки в написанні однакових пунктів, різні правила заповнення, різний порядок слів в назві, наявність чи відсутність додаткових кодів, марок та одиниць виміру ТМЦ).
3. Виправлення помилок та розбіжностей структуризації класифікатора, а саме наявності чи відсутності різних груп чи об'єктів.

1. Математична модель

Дамо деякі визначення.

$\overline{E} = [E_i]_{i=1}^{n_e}, n_e = \left\|\overline{E}\right\|$ — множина об'єктів E класифікатора, де n_e — кількість об'єктів.

$\overline{K} = \left[\overline{E}_i\right]_{i=1}^{n_k}, n_k = \left\|\overline{K}\right\|$ — ієрархічний класифікатор, заданий як множина множин об'єктів.

$\overline{E}^0 = \bigcup_{i=1}^{n_k} \overline{E}_i$ — множина нульового рівня класифікації, представляє собою множину всіх об'єктів класифікатора.

$\forall \overline{E}^l, l > 0\ \exists! \overline{E}^{l-1} : \overline{E}^l \subseteq \overline{E}^{l-1}$ — множина l рівня класифікації, вкладена рівно в одну множину $l-1$ рівня.

За визначенням класифікатора, об'єкт може бути вкладений тільки в одну множину на кожному з рівнів, причому множини повинні бути послідовно вкладеними. Тому можна побудувати множину всіх послідовно вкладених множин, що містять об'єкт:

$$\overline{B}\left(E^m\right) = \left[\overline{E}^l\right]_{l=1}^{m}, \forall l \in [1;m]\ \exists \overline{E}^{l-1} : \overline{E}^l \subseteq \overline{E}^{l-1}, E^m \in \overline{E}^l.$$

Тут m — рівень об'єкта, це рівень найменшої множини, що містить заданий об'єкт E^m. Якщо розглядати класифікатор як дерево, то $\overline{B}(E)$ — це гілка дерева, листовим елементом якої є заданий об'єкт E.

1.1. **Постановка задачі.** Нехай задана множина ієрархічних класифікаторів:

$$\overline{F} = \left[\overline{K}_i\right]_{i=1}^{n_F}, n_F = \left\|\overline{F}\right\|,$$

причому множини об'єктів класифікаторів можуть перетинатися

$$\overline{I} = \bigcup_{j=1,k=1}^{n_F,n_F} \overline{E}_j^0 \bigcap \overline{E}_k^0, \overline{E}_j^0 \in \overline{K}_j, \overline{E}_k^0 \in \overline{K}_k, \left\|\overline{I}\right\| \geq 0.$$

Задача полягає у побудові зведеного класифікатору $\overline{K}^F = \bigcup_{k=1}^{n_F} \overline{E}_k^0$, що містить елементи всіх заданих класифікаторів. Головною задачею, що необхідно вирішити, є зменшення дублювання об'єктів — $\left\|\overline{I}\right\| \to 0$.

1.2. **Нечітка відповідність між множинами.** Нехай задана попарна нечітка відповідність між об'єктами класифікаторів $s\left(E_i, E_j\right) \in [0;1]$.

Тоді відповідність між об'єктом та множиною є максимальна відповідність між об'єктом E_i та об'єктами множини $\overline{E}_j$:

$$d\left(E_i, \overline{E}_j\right) = \max_{k=1}^{n}\left(s\left(E_i, E_j\right)\right), n = \left\|\overline{E}_j\right\|.$$

Звідси можна визначити сумарну відповідність між множинами як суму відповідностей об'єктів множини $\overline{E}_i$ до множини $\overline{E}_j$:

$$\hat{d}\left(\overline{E}_i, \overline{E}_j\right) = \sum_{k=1}^{n_i} d\left(E_k, \overline{E}_j\right), E_k \in \overline{E}_i, n_i = \left\|\overline{E}_i\right\|.$$

Визначимо односторонню нечітку відповідність між множинами $\overline{E}_i$ та $\overline{E}_j$ як суму відповідностей об'єктів множини $\overline{E}_i$ до множини $\overline{E}_j$, поділена на кількість елементів множини $\overline{E}_i$:

$$l\left(\overline{E}_i, \overline{E}_j\right) = \frac{\hat{d}\left(\overline{E}_i, \overline{E}_j\right)}{\left\|\overline{E}_i\right\|}.$$

Нехай задана множина множин $M = \left[\overline{E}_i\right]_{i=1}^{m}, m \geq 2$. Дамо наступні визначення.

$\tilde{d}\left(\overline{E}_i, M\right) = \sum\limits_{j=1, j\neq i}^{m} \hat{d}\left(\overline{E}_i, \overline{E}_j\right)$ — сумарна відповідність множини $\overline{E}_i$ до всіх інших множин з M.

$\tilde{n}\left(\overline{E}_i, M\right) = \sum\limits_{j=1, j\neq i}^{m} \left\|\overline{E}_j\right\|$ — сумарна кількість об'єктів всіх множин з M, крім множини $\overline{E}_i$.

Тоді визначимо групову нечітку відповідність між множинами з множини M як суму парних відповідностей між об'єктами кожної з множин множини M, поділену на суму кількості елементів кожної з множин множини M :

$$g(M) = \frac{\sum\limits_{i=1}^{m} \tilde{d}\left(\overline{E}_i, M\right)}{\sum\limits_{i=1}^{m} \tilde{n}\left(\overline{E}_i, M\right)}, m = \left\|\overline{M}\right\|, \overline{E}_i \in M.$$

Нехай $p \in [0;1]$ - заданий поріг. Задамо наступні предикати попарної відповідності множин:

1. $P_I\left(p, \overline{E}_i, \overline{E}_j\right) : l\left(\overline{E}_i, \overline{E}_j\right) \geq p$ — вкладеність множини $\overline{E}_i$ в множину $\overline{E}_j$ зі ступенем $l\left(\overline{E}_i, \overline{E}_j\right)$.
2. $P_L\left(p, \overline{E}_i, \overline{E}_j\right) : P_I\left(p, \overline{E}_i, \overline{E}_j\right) \wedge \neg P_I\left(p, \overline{E}_j, \overline{E}_i\right)$ — точна вкладеність множини $\overline{E}_i$ в множину $\overline{E}_j$ зі ступенем $l\left(\overline{E}_i, \overline{E}_j\right)$.
3. $P_G(p, M) : g(M) \geq p$ — групова відповідність між множинами множини M зі ступенем $r(M)$.
4. $P_R(p, M) : \bigwedge\limits_{i=1}^{m} \bigwedge\limits_{j=1, j\neq i}^{m} P_I\left(p, \overline{E}_i, \overline{E}_j\right)$ — точна групова відповідність між множинами множини M.

Легко бачити, що $P_R(p, M) \Rightarrow P_G(p, M)$, але обернене невірно, звідси може виникнути ситуація, коли для пари множин одночасно виконуються предикати P_L та P_G. Тому для виявлення множин для об'єднання необхідно використовувати більш сильний предикат P_R.

1.3. **Об'єднання множин.** Розглянемо задачу об'єднання множин l рівня класифікації, де $l > 0$.

Визначимо оператор, що для заданої множини $\overline{E}^{l-1}$ знаходить всі множини l рівня класифікації, для яких виконується предикат точної групової відповідності множин та які вкладені в задану множину:

$$F_M^U\left(\overline{E}^{l-1}\right) = \left[\overline{M}_i^l\right]_{i=1}^{m}, \overline{M}_i^l = \left[\overline{E}_j^l\right], \overline{E}_j^l \in \overline{E}^{l-1}, P_R\left(p, \overline{M}_i^l\right),$$

причому він вибирає ці множини так, щоб вони не перетиналися:

$$\forall i \in [1;m] \, \forall j \in [1;m], j \neq i : \overline{M}_i^l \bigcap \overline{M}_j^l = \emptyset.$$

Також визначимо оператор об'єднання множин, який для множини $\overline{M}_i^l$ будує нову множину l рівня класифікації шляхом об'єднання множин, з яких вона складається:

$$F_U\left(\overline{M}_i^l\right) = \bigcup_{j=1}^{m_i} \overline{E}_j^l = \bigcup_{j=1}^{m_i} \left[\overline{E}_k^{l+1}\right]_{k=1}^{n_j}.$$

1.4. **Вкладення множин.** Розглянемо задачу вкладення множин l рівня класифікації, де $l > 0$.

Визначимо оператор, що для заданої множини $\overline{E}^{l-1}$ знаходить всі пари множин $\overline{E_a}^l$ та $\overline{E_b}^l$ l рівня класифікації, для яких виконується оператор точної вкладеності множини та які вкладені в задану множину:

$$F_M^I\left(\overline{E}^{l-1}\right) = \left[\left(\overline{E_{a_i}}^l, \overline{E_{b_i}}^l\right)\right]_{i=1}^{m}, \overline{E_{a_i}}^l \in \overline{E}^{l-1}, \overline{E_{b_i}}^l \in \overline{E}^{l-1}, P_L\left(p, \overline{E_{a_i}}^l, \overline{E_{b_i}}^l\right),$$

причому він вибирає ці пари множин так, щоб кожна множина зустрічалася в результуючій множині тільки один раз:

$$\bigcap_{i=1}^{m_l} \overline{E_{a_i}}^l \cap \bigcap_{i=1}^{m_l} \overline{E_{b_i}}^l = \emptyset.$$

Також задамо оператор об'єднання даних множин, що проводить обробку вкладення шляхом або об'єднання двох множин в одну або вкладення однієї множини в іншу, в залежності від предметної області і конкретної пари множин:

$$F_I\left(\overline{E_a}^l, \overline{E_b}^l\right) = \overline{E}^{l\prime} = \begin{cases} \left[\overline{E_{a_k}}^{l+1}\right] \bigcup \left[\overline{E_{b_k}}^{l+1}\right], \\ \left[\overline{E_a}^l\right] \bigcup \left[\overline{E_{b_k}}^{l+1}\right]. \end{cases}$$

1.5. **Алгоритм побудови зведеного класифікатора.** Таким чином, можна навести наступний алгоритм побудови зведеного класифікатора:

1. Побудувати об'єднану множину нульового рівня класифікації, що містить об'єднання множин першого рівня всіх класифікаторів:
$$\overline{E}^{0\prime} = \bigcup_{k=1}^{n_F} \left[\overline{E}^1\right]_k.$$
2. Провести об'єднання множин, для кожної множини $\overline{E}^{l-1}$, починаючи з $\overline{E}^0$:
   a) Застосувати оператор пошуку множин, що необхідно об'єднати, до множини $\overline{E}^{l-1}$:
   $$\left[\overline{M}^l\right] = F_M^U\left(\overline{E}^{l-1}\right).$$
   b) Застосувати оператор побудови об'єднаних множин F_U до кожної отриманої множини:
   $$\overline{E}^{l\prime} = F_U\left(\overline{M}^l\right).$$
   c) Провести об'єднання множин для кожної отриманої множини $\overline{E}^{l\prime}$.

3. Провести обробку вкладення множин, для кожної множини $\overline{E}^{l-1'}, l > 0$:
   a) Застосувати оператор пошуку множин, для яких необхідно провести обробку вкладення, до множини $\overline{E}^{l-1'}$:
   $$\left[\left(\overline{Ea}^{l'}, \overline{Eb}^{l'}\right)\right] = F_M^I\left(\overline{E}^{l-1'}\right).$$
   b) Застосувати оператор побудови об'єднаних множин F_I до кожної отриманої множини:
   $$\overline{E}^{l''} = F_I\left(\overline{Ea}^{l'}, \overline{Eb}^{l'}\right).$$
   c) Провести обробку вкладення кожної отриманої множини $\overline{E}^{l''}$.

2. Реалізація побудови зведеного ієрархічного реєстру

Пропонується консолідацію, очистку та узгодження даних при створенні централізованої системи НДІ здійснювати за наступною схемою:

1. Консолідація даних з окремих класифікаторів ТМЦ філій корпорації — уточнення структури та збір даних до єдиного місця зберігання; завантаження даних з баз даних класифікаторів підрозділів до об'єднаної БД.
2. Стандартизація даних — приведення значень атрибутів до узгодженого формату зберігання, форматування текстових даних: переведення в верхній регістр, видалення зайвих пробілів та символів - «паразитів».
3. Ідентифікація записів — знаходження нечіткої подібності між записами всередині одного класифікатору та між різними класифікаторами, використовуючи відстань Левенштейна та метод Q-грам [4, 5]. Головною характеристикою запису при знаходженні ступеню подібності до інших записів є його назва та набір текстових атрибутів.
4. Ідентифікація груп — знаходження нечіткої подібності між різними групами записів класифікаторів, використовуючи ступінь подібності між записами груп. Так як одні й ті ж групи в процесі побудови класифікаторів в різних філіях корпорації могли мати різні назви, то головною характеристикою групи при ступеню подібності до інших груп є не її назва, а множина всіх вкладених підгруп та записів.
5. Об'єднання груп з різних класифікаторів, що мають ступінь подібності більше заданого порогу. Якщо групи визначені, як подібні, то множини об'єктів цих груп об'єднуються в одну групу, якщо множина елементів однієї групи визначена як вкладена в множину елементів іншої групи, то перша група переноситься в другу групу як підлеглий елемент, разом зі всіма своїми елементами.
6. Оптимізація даних. Після об'єднання даних проводиться аналіз класифікатора на помилки та неоптимальну організацію даних в ієрархії. Даний етап проводиться за участю експерта.

2.1. **Консолідація даних.** Консолідація даних включає в себе уніфікацію структури даних — завантаження даних з різних джерел в БД єдиної структури.

Для завантаження даних з різних джерел необхідно визначити структуру збереження класифікаторів для БД кожної філії. Для цього вирішуються наступні задачі:

1. Визначення переліку всіх множин та об'єктів кожної з БД філій.
2. Визначення переліку атрибутів множини та об'єкту кожної з БД філій.
3. Визначення значень атрибутів кожної множини та атрибутів кожного об'єкту.
4. Однозначне віднесення кожної множини та кожного об'єкту до тієї чи іншої множини, в рамках даних БД однієї філії.
5. Визначення переліку атрибутів множини та об'єкту для об'єднаного класифікатора.
6. Визначення відображення атрибутів множини та об'єкту кожної з БД філій до атрибутів множини та об'єкту об'єднаного класифікатора.

Дані задачі можливо автоматизувати, застосовуючи систему правил для кожної БД філії. У випадку, коли неможливо провести автоматичну обробку, необхідна участь експерта у вирішенні задачі.

З точки зору математичної моделі даний етап є реалізацією першого етапу алгоритму — побудови об'єднаної множини нульового рівня класифікації.

2.2. **Стандартизація даних.** Стандартизація даних включає в себе виділення елементів атрибутів та уніфікацію — приведення представлення атрибутів об'єктів до єдиного формату.

Елементи атрибутів об'єктів можуть містити помилки, але за можливістю нечіткої обробки можна виділити наступні класи елементів:

1. Нечутливі до змін елементи — елементи, до яких можна застосувати нечітку ідентифікацію. До них відносяться терміни, що описують об'єкт, наприклад, «Металлорукав», «Полоса медная», «Лист свинц.», «діаметр».
2. Чутливі до змін елементи — елементи, до яких не можна застосовувати нечітку ідентифікацію внаслідок того, що зміна будь-якого символу призводить до повної зміни значення. До них відносяться:
   a) Спеціальні ідентифікатори – скорочення, абревіатури, коди, одиниці виміру, наприклад, «мм», «РП-Ц-А-75», «ГОСТ».
   b) Числові ідентифікатори – розміри, діаметри тощо.

Для виділення елементів атрибутів об'єктів застосовуються системи правил. Так як дані в різних БД філій можуть сильно відрізнятися за структурою та складом атрибутів, то застосовуються різні системи правил для даних, що були завантажені з різних БД філій.

Після виділення типів елементів проводиться уніфікація даних, що полягає у перекодуванні символів рядків (наприклад, переведення в верхній регістр, перекодування символів) та елементів атрибутів (наприклад, з використанням таблиць відповідності).

Результатом даного етапу є множина об'єктів класифікатора, приведена до єдиного стандарту.

2.3. **Ідентифікація записів.** Ідентифікація записів полягає в знаходженні відповідності між записами про об'єкти класифікатора в БД. Для цього для кожного запису в БД знаходяться записи, що точно співпадають з даним записом та ті, що мають ступінь подібності більше заданого порогу.

Якщо розглядати об'єкти як множини елементів атрибутів, то ступінь відповідності між об'єктами класифікатора E_i та E_j можна визначити як групову нечітку відповідність множини множин $[E_i, E_j]$:

$$s\left(E_i, E_j\right) = g\left(\left[E_i, E_j\right]\right).$$

Ідентифікація об'єктів прямим перебором має велику складність, тому використовується підготовка елементів — нечітка ідентифікація елементів та збереження інформації про результати ідентифікації в БД. Для нечутливих до змін елементів ступінь відповідності між елементами може бути побудована на основі відстані Левенштейна $d_l\left(A_i^M, A_j^M\right)$ чи на основі різниці Q-грам, зокрема, на основі біграм $d_q\left(A_i^M, A_j^M\right), q = 2$.

Нехай задані пороги відповідності $p^A \in [0;1]$ та $p^E \in [0;1]$.

Для підготовки елементи зберігаються в БД та прив'язується до об'єкту, до якого вони відносяться. Якщо елемент вже існує в БД, він повторно не додається, а тільки прив'язується до об'єкту. Крім того, між елементами, для яких ступінь відповідності більше заданого порогу p^A, зберігаються парні зв'язки із зазначенням ступеню відповідності.

Після підготовки проводиться ідентифікація об'єктів, що полягає в знаходженні та збереженні об'єктів, що подібні до заданого. Для кожного заданого об'єкту E_i виконуються наступні операції:

1. Пошук об'єктів — проводиться пошук записів, що можуть бути подібними до заданого. Проводиться з використанням підготованих елементів:
   a) Вибираються всі елементи, прив'язані до заданого об'єкту.
   b) Вибираються всі об'єкти, до яких прив'язані відібрані елементи.
2. Ідентифікація об'єктів — проводиться розрахунок групової нечіткої відповідності для кожної пари заданого об'єкту і відібраного об'єкту E_j.
3. Збереження результатів ідентифікації — для кожного відібраного об'єкту, для якого виконується предикат точної групової відповідності із заданим об'єктом $P_R\left(p^E, [E_i, E_j]\right)$, зберігається зв'язок між цим об'єктом та заданим об'єктом, із зазначенням ступеню відповідності.

Таким чином, результатом даного етапу є збережена інформація про парну нечітку відповідність між об'єктами класифікатора.

2.4. **Ідентифікація груп.** Ідентифікація груп полягає у знаходженні та збереженні односторонніх нечітких відповідностей між множинами класифікатору із заданим порогом $p^M \in [0;1]$.

Для кожної множини $\overline{E}_i^l$ виконуються наступні операції:

1. Пошук множин — для зменшення алгоритмічної складності проводиться попередній пошук множин, що можуть бути подібними до заданої:

a) Вибираються всі об'єкти заданої множини.
b) Вибираються всі об'єкти, прив'язані до вибраних на попередньому кроці
c) З множини $\overline{E}^{l-1}$, в яку вкладена множина $\overline{E}_i^l$, вибираються множини $\left[\overline{E}_j^l\right]_{j=1}^n$, до яких прив'язані об'єкти, вибрані на попередньому кроці:

$$\overline{E}_i^l \in \overline{E}^{l-1} \wedge \forall j \in [1;n] : \overline{E}_j^l \in \overline{E}^{l-1}.$$

2. Ідентифікація множин — для кожної відібраної множини $\overline{E}_j$ обраховується одностороння нечітка відповідність $l\left(\overline{E}_j, \overline{E}_i\right)$.
3. Збереження множин — для кожної відібраної множини, для якої виконується предикат вкладеності множин $P_I\left(p^M, \overline{E}_i, \overline{E}_j\right)$, зберігається зв'язок між цією множиною та заданою множиною, із зазначенням ступеню відповідності.

Таким чином, результатом даного етапу є збережена інформація про вкладеність множин класифікатора.

2.5. **Об'єднання груп.** Об'єднання груп полягає в побудові множин, для яких виконується предикат точної групової відповідності та об'єднання цих множин.

Для цього задається рівень $l = 1$ та вибирається множина верхнього рівня $\overline{E}^{l-1} = \overline{E}^0$, для якої виконуються наступні операції:

1. Побудова множини $M = \left[\overline{E}_i^l\right]_{i=1}^m$:
   a) Вибрана множина $\overline{E}_k^l$ утворює множину M.
   b) До множини M додається множина $\overline{E}_m^l$, що належить до множини $\overline{E}^{l-1}$ та має прямі та зворотні зв'язки зі всіма множинами з M, тобто виконується предикат точної групової відповідності $P_R\left(p^M, M \bigcup \left[\overline{E}_m^l\right]\right)$. Якщо таких множин декілька, то вибирається та, у якої сума ступенів відповідності всіх зв'язків максимальна.
   c) Операція 1b повторюється до тих пір, поки знаходяться множини $\overline{E}_m^l$.

2. Об'єднання множин $\overline{E}_i^l$ до множини $\overline{E}_r^l = \bigcup_{i=1}^m \overline{E}_i^l$:
   a) Створюється множина $\overline{E}_r^l$. Атрибути множини $\overline{E}_r^l$ формуються на основі атрибутів множин $\overline{E}_i^l$. Зокрема, проводиться конкатенація назв множин, якщо вони відрізняються в різних множин.
   b) Всі множини та об'єкти, що позначені як вкладені в $\overline{E}_i^l$, позначаються як вкладені в $\overline{E}_r^l$.
   c) Всі множини $\overline{E}_i^l$ видаляються.

3. Повторюються операції 1 та 2 до тих пір, поки на заданому рівні є множини, для яких є пов'язані множини.

4. Рекурсивно виконуються всі операції для всіх підмножин множини $\overline{E}^{l-1}$.

З точки зору математичної моделі даний етап є реалізацією другого етапу алгоритму – об'єднання множин, починаючи з верхнього рівня.

2.6. **Оптимізація даних.** Після проведення зведення даних в єдину БД необхідно провести аналіз дерева на повтори та неоптимальну організацію даних за участю експертів.

Оптимізація даних включає наступні операції:

1. Ідентифікація записів про об'єкти класифікатора. Полягає в об'єднанні записів про об'єкти класифікатора, що визнані записами про один об'єкт. Для цього експерту надається для обробки перелік пар пов'язаних об'єктів та надається можливість об'єднати об'єкти або відмінити об'єднання.
2. Обробка атрибутів множин. Після об'єднання декількох множин в одну атрибути об'єднаної множини формуються автоматично, але отримані атрибути потребують ручної обробки. Для цього експерту надається перелік об'єднаних множин та надається можливість ручного редагування атрибутів обраної множини.
3. Обробка вкладення множин. Для пар множин, для яких виконується предикат вкладеності множини, проводиться вкладення або об'єднання множин залежить від множин та предметної області. В разі об'єднання множин для об'єднаної множини проводиться ручна обробка її атрибутів.

Таким чином, система надає експерту перелік варіантів для прийняття рішення щодо кожного з пунктів та забезпечує виконання прийнятих рішень. Після проведення даних операцій можна продовжувати ведення елементів класифікатора, що включає редагування атрибутів обраного об'єкту чи групи, та переміщення об'єкту чи групи до іншої групи.

Висновки

В статті запропоновано алгоритм побудови зведеного ієрархічного реєстру ТМЦ. Особливістю даного алгоритму є те, що кожна група класифікатора розглядається як множина об'єктів, які відносяться до групи. Це дозволяє відкинути розбіжності в значеннях атрибутів групи, таких, як назва групи, при завантаженні даних із різних класифікаторів.

Даний алгоритм було реалізовано для побудови єдиного класифікатора ТМЦ корпорації, сформованого на основі класифікаторів чотирьох філій. Середній об'єм класифікаторів філій – 100000 об'єктів. Класифікатори ТМЦ філій мали схожу структуру класифікації, але в процесі незалежного використання в кожній з філій в структурі класифікаторів з'явилися розбіжності. В результаті проведеного аналізу були вибрані наступні пороги $p^A = 0.75$, $p^E = 0.75$, $p^M = 0.45$, що дозволило об'єднати близько 68% груп класифікаторів.

Одним з напрямків подальшого розвитку роботи є розробка рекомендацій щодо підбору порогів нечіткої відповідності для елементів атрибутів об'єктів, об'єктів та множин.

Список литературы

1. *ДСТУ 1.10:2005* Національна стандартизація. Правила розроблення, побудови, викладання, оформлення, ведення національних класифікаторів.
2. *Rahm E, Do H.H.* Data Cleaning: Problems and Current Approaches // IEEE Techn. Bulletin on Data Engineering, Dec. – 2000.
3. *Kimball R., Caserta J.* The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data. Wiley 2004.
4. *Орлов Д.* Подсистема сопоставления записей в хранилище данных. http://www.olap.ru/basic/CompareLog_dw.asp.
5. *Гула А.Ю., Игнатенко А.П., Перечинский И.А.* Применение методов интеллектуальной обработки в задачах очистки хранилища данных // Сб. трудов конф. Системы поддержки принятия решений. Теория и практика. – Киев, 2007, С. 145 – 148.

Статья поступила в редакцию 30.04.2008