ДОДАТОК А

Програмний код

Процедура з оцінювання якості персоналізації пошуку

```
FROM #vector_page
WHERE page_id IN (34, 42/*, 45*/)and page_id <> 504
GROUP BY word_id
UNION


SELECT 1, word_id, power = 50
FROM #vector_page WHERE page_id = 504
GROUP BY word_id

INSERT #vector_user (user_id, word_id, power)
SELECT 4, word_id, MAX(power)
FROM #vector_page
WHERE page_id IN (8, 20, 22, 44, 41, 43, 46)and page_id <> 504
GROUP BY word_id
UNION
SELECT 1, word_id, power = 50
FROM #vector_page WHERE page_id = 504
GROUP BY word_id


IF OBJECT_ID('tempdb..#euklidDistTable') IS NOT NULL DROP TABLE
#euklidDistTable
CREATE TABLE #euklidDistTable ([object_id] int, [avgCluster] int,
[EuklidDist] float, PRIMARY KEY ([object_id], [avgCluster]))

PRINT 'Расчёт евклидово расстояние относительно центроидов'
INSERT #euklidDistTable ([object_id], [avgCluster],   [EuklidDist])
SELECT            GOCT.[page_id],        avgCluster =
      aVT.[user_id],       EuklidDist    =
      POWER(SUM(POWER((GOCT.[power]      -
aVT.[power]), 2)),0.5)
FROM #vector_page GOCT
INNER JOIN
(
      select [user_id], [word_id], [power]
from #vector_user ) AS aVT
ON aVT.[word_id] = GOCT.[word_id]
GROUP BY GOCT.[page_id], aVT.[user_id]
ORDER BY GOCT.[page_id], aVT.[user_id]OPTION (RECOMPILE);
PRINT 'Расчёт евклидово расстояние относительно центроидов выполнено'

  для каждого объекта находим кластер с минимальным растоянием к центру.
IF OBJECT_ID('tempdb..#euklidDistTableMin') IS NOT NULL DROP TABLE
#euklidDistTableMin
CREATE TABLE #euklidDistTableMin ([object_id] int not null, [avgCluster]
int, [EuklidDistMin] float)
INSERT #euklidDistTableMin ([object_id], [avgCluster], [EuklidDistMin])
```

```
select top(1) with ties [object_id],
[avgCluster], [EuklidDist] FROM
#euklidDistTable eDT
order by row_number() over (partition by [object_id] order by EuklidDist
ASC)

IF OBJECT_ID('tempdb..#GlobalObjectClusterTable1') IS NOT NULL DROP TABLE
#GlobalObjectClusterTable1 Create table
#GlobalObjectClusterTable1([object_id] int not null, [word] varchar(50),
[power] float, [cluster] int)
  Загружаем ИР
INSERT INTO #GlobalObjectClusterTable1 ([object_id], [word], [power],
[cluster])
SELECT page_id, CAST(word_id as varchar(10)), power, 0 FROM #vector_page
  Загружаем ИП
--INSERT INTO #GlobalObjectClusterTable1 ([object_id], [word], [power],
[cluster])
--SELECT -[user_id], CAST(word_id as varchar(10)), power, [user_id] FROM
#vector_user


UPDATE GOCT SET GOCT.[cluster] = C.[avgCluster]
FROM
#GlobalObjectClusterTable1 GOCT
INNER JOIN
#euklidDistTableMin C
ON GOCT.[object_id] = C.[object_id]

IF EXISTS (SELECT NULL FROM #GlobalObjectClusterTable1 WHERE [cluster] = 0)
BEGIN
     RAISERROR('Имеются не кластеризованные объекты',11,1);
END

PRINT
'/**************************************************************************
**************************
**\'
PRINT '|*********************Результат предварительной
кластеризации*****************************************|'
PRINT
'\**************************************************************************
**************************
**/'
IF OBJECT_ID('tempdb..#ClusterContent1') IS NOT NULL DROP TABLE
#ClusterContent1
     CREATE TABLE #ClusterContent1 ([cluster] int NOT NULL PRIMARY KEY,
[cluster_cntOBJ] int, [OBJS] varchar(max))
     INSERT #ClusterContent1 ([cluster], [cluster_cntOBJ])
     select [cluster], COUNT(DISTINCT [object_id])

          FROM #GlobalObjectClusterTable1
          GROUP BY [cluster]
          ORDER BY [cluster]

     UPDATE CC SET
   CC.OBJS =    REPLACE(
                     REPLACE(
                          CAST(
                                   (
                                             Select
```

```
                                    distinct
convert(varchar(1000),
[object_id]) AS 'Itm'



                                  ),
         FROM #GlobalObjectClusterTable1 GOCT
         WHERE GOCT.cluster = CC.cluster
         FOR XML PATH('')
)AS varchar(MAX)
        --//получаем строку
вида
<Itm>ids1</Itm><Itm>Ids2</Itm><Itm>Ids3</Itm>...
         '<Itm>',''),
       '</Itm>', ';')
      from
           #ClusterContent1 CC

    --// удаляем последнюю
             точку-запятую
   UPDATE CC SET CC.OBJS =
                      CASE
                                              WHEN
                                                  SUBSTRING(CC.OBJS,
                                                  LEN(CC.OBJS), 1)         =
';' THEN LEFT(CC.OBJS, LEN(CC.OBJS) -1)
                                              ELSE CC.OBJS
                                         END
      FROM #ClusterContent1 CC

      declare @curs2_cluster int;
      declare @curs2_cntOBJ int;
      declare @curs2_OBJS varchar(max);
      DECLARE curs2 cursor LOCAL FAST_FORWARD for
           SELECT [cluster], [cluster_cntOBJ], [OBJS]
      FROM #ClusterContent1 OPEN curs2
      FETCH curs2 INTO @curs2_cluster, @curs2_cntOBJ, @curs2_OBJS
      WHILE @@FETCH_STATUS = 0
      BEGIN
           PRINT 'В кластере с номером: ' +
           CONVERT(varchar(100),@curs2_cluster) + ' содержится:

           '
  CONVERT(varchar(100), @curs2_cntOBJ) + ' объектов:';
           PRINT '{';
           PRINT ' ' +
           @curs2_OBJS;
           PRINT '}';
           FETCH curs2 INTO @curs2_cluster, @curs2_cntOBJ, @curs2_OBJS;
      END
      CLOSE curs2;
      DEALLOCATE curs2;

PRINT '==========================================Конец
отчёта================================================';
```

```
IF OBJECT_ID('tempdb..#GlobalObjectClusterTableMAIN') IS NOT NULL DROP
TABLE #GlobalObjectClusterTableMAIN
CREATE TABLE #GlobalObjectClusterTableMAIN
([object_id] int NOT NULL, [word] varchar(50) NOT NULL, [power] float NOT
NULL, [cluster] int NOT NULL, PRIMARY KEY ([object_id], [word]))
INSERT INTO #GlobalObjectClusterTableMAIN ([object_id],
[word], [power], [cluster]) SELECT [object_id], [word],
[power], [cluster] FROM #GlobalObjectClusterTable1


IF OBJECT_ID('tempdb..#GlobalObjectClusterTableMAINFINAL') IS NOT NULL DROP
TABLE #GlobalObjectClusterTableMAINFINAL
CREATE TABLE #GlobalObjectClusterTableMAINFINAL
([object_id] int NOT NULL, [word] varchar(50) NOT NULL, [power] float NOT
NULL, [cluster] int NOT NULL, PRIMARY KEY ([object_id], [word]))
INSERT INTO #GlobalObjectClusterTableMAINFINAL ([object_id],
[word], [power], [cluster]) exec
InternetDB2.dbo.az_clustering_step3
PRINT
'/*************************************************************************
***************************
**\'
PRINT '|*********************Результат
кластеризации*******************************************|'
PRINT
'\*************************************************************************
***************************
**/'

IF OBJECT_ID('tempdb..#ClusterContent') IS NOT NULL DROP TABLE
#ClusterContent
     CREATE TABLE #ClusterContent ([cluster] int NOT NULL PRIMARY KEY,
[cluster_cntOBJ] int, [OBJS] varchar(max))
     INSERT #ClusterContent ([cluster], [cluster_cntOBJ])
     select [cluster], COUNT(DISTINCT [object_id])
          FROM #GlobalObjectClusterTableMAINFINAL
          GROUP BY [cluster]
          ORDER BY [cluster]

     UPDATE CC SET
   CC.OBJS =    REPLACE(
                     REPLACE(
                          CAST(
                                    (
                                             select distinct
                                                    convert(varchar(1000),
[object_id]) AS 'Itm'
                                             FROM
                                             #GlobalObjectClusterTableMAINFIN
                                             AL GOCT
                                             WHERE GOCT.cluster = CC.cluster
                                             FOR XML PATH('')
                                    )AS varchar(MAX)
                               ), --//получаем строку вида
<Itm>ids1</Itm><Itm>Ids2</Itm><Itm>Ids3</Itm>...
        '<Itm>',''),
       '</Itm>', ';')
     from
          #ClusterContent CC
```

```
    --// удаляем последнюю
              точку-запятую
  UPDATE CC SET CC.OBJS =
                       CASE
 WHEN SUBSTRING(CC.OBJS, LEN(CC.OBJS), 1) =
';' THEN LEFT(CC.OBJS, LEN(CC.OBJS) -1)
                                              ELSE CC.OBJS
                                         END
     FROM #ClusterContent CC
     declare @curs3_cluster int;
     declare @curs3_cntOBJ int;
     declare @curs3_OBJS varchar(max);
     DECLARE curs3 cursor LOCAL FAST_FORWARD for
          SELECT [cluster], [cluster_cntOBJ], [OBJS]
     FROM #ClusterContent OPEN curs3
     FETCH curs3 INTO @curs3_cluster, @curs3_cntOBJ, @curs3_OBJS
     WHILE @@FETCH_STATUS = 0
     BEGIN
          PRINT 'В кластере с номером: ' +
          CONVERT(varchar(100),@curs3_cluster) + ' содержится:

          '
 CONVERT(varchar(100), @curs3_cntOBJ) + ' объектов:';
          PRINT '{';
          PRINT ' ' +
          @curs3_OBJS;
          PRINT '}';
          FETCH curs3 INTO @curs3_cluster, @curs3_cntOBJ, @curs3_OBJS;
     END
     CLOSE curs3;
     DEALLOCATE curs3;
```

ДОДАТОК Б

Слайди презентації

Харківський національний університет радіоелектроніки

Атестаційна робота магістра

Дослідження методів кластерізації ресурсів інформаційних мереж

Виконала
ст. гр. ІПЗмзд-17-1
Баткова Т.В.

Науковий керівник
проф. Шубін І.Ю.

1

Мета роботи

- Методи кластеризації динамічних об'єктів, таких як IP, недостатньо розроблені й, крім того, мало хто з дослідників розглядав ідею узагальненого показу об'єктів різної природи, що мають подібні властивості.
- Метою є застосування методів кластерного аналізу для класифікації ІК та IP задля персоналізації інформаційного пошуку і підвищення

Об'єкт дослідження

методи персоналізації Інтернет- пошуку, засновані на вивченні й класифікації ІК та ІР за допомогою кластерного аналізу.

- Перехід від вербальної до числової репрезентації координат відбувається за рахунок позиційного кодування термінів і підрахунку числа їх входжень у текст пошукових запитів або текстовий контент статичних компонентів Dom-моделі ІР.
- Розроблений набір програмних модулів (програмна система) для спостереження за активністю ІК і одержання текстового змісту ІР з їх обліком .

3

Класифікація ІК по СД ознаках або ІР за структурою

- **Графіки ваг слів для ІК, розділених по СД ознаках**

4

Числові координати

- розташовані в характеристичному векторі в порядку, відповідному до лексикографічного порядку проходження відповідних термінів у словнику Vu.
- Перехід від вербального до числового представлення результатів відбувається за рахунок позиційного кодування термінів і підрахунку числа їх входжень у запити пошукової історії ІК.
- Перенесено математичний опис абстрактних об'єктів в область дослідження пошукових запитів ІК

5

- У довільний момент часу $t_k \in T$, визначений ІК можна представити характеристичним вектором наступного виду:

$$u_i(t_k) = (u_{i,1}(t_k), \ldots, u_{i,j}(t_k), \ldots, u_{i,nof(Vu)}(t_k)),$$

6

Схема поетапного процесу виконання завдань класифікації

- Етап 1. Настановний.
- Етап 2. Постановочний
- Етап 3. Інформаційний
- Етап 4. Апріорно математико-постановочний
- Етап 5. Розвідницький аналіз.
- Етап 6. Апостеріорний математико-постановочний
- Етап 7. Обчислювальний.
- Етап 8. Підсумковий.

В залежності від результатів цього аналізу, або формуються остаточні наукові чи прикладні висновки, або даються уточнення й доповнення до завдання й відбувається

7

Семантичний аналіз

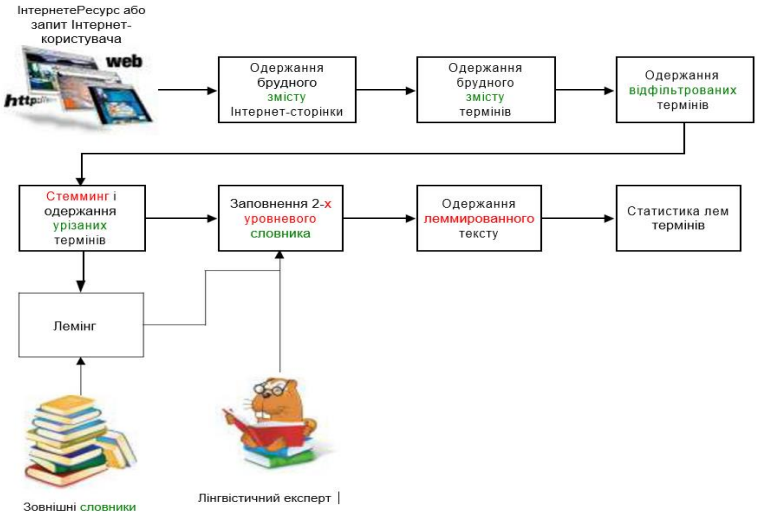
дозволяє виявляти незв'язність слів і речень у тексті, хоча вони й можуть бути погоджені граматично.

дозволяє визначати метафори, переносні значення, дійсний зміст співзвучних слів залежно від контексту, і т.д.

- Прості способи семантичного аналізу дозволяють класифікувати текст, виділяти емоційне забарвлення тексту (за допомогою виявлення певних слів і аналізу словосполучень, що містять метафори і алегорії) і його тему (по синтаксичних ознаках і кількості повторюваних слів у реченнях).
- Зокрема, за допомогою семантичного аналізу відбувається видача контекстної реклами на багатьох сайтах і в пошукових системах.
- Сторінка, запропонована користувачеві, досліджується на предмет наявності повторюваних ключових слів, після чого автоматизований генератор реклами видає пов'язану зі знайденими ключовими словами вибірку.

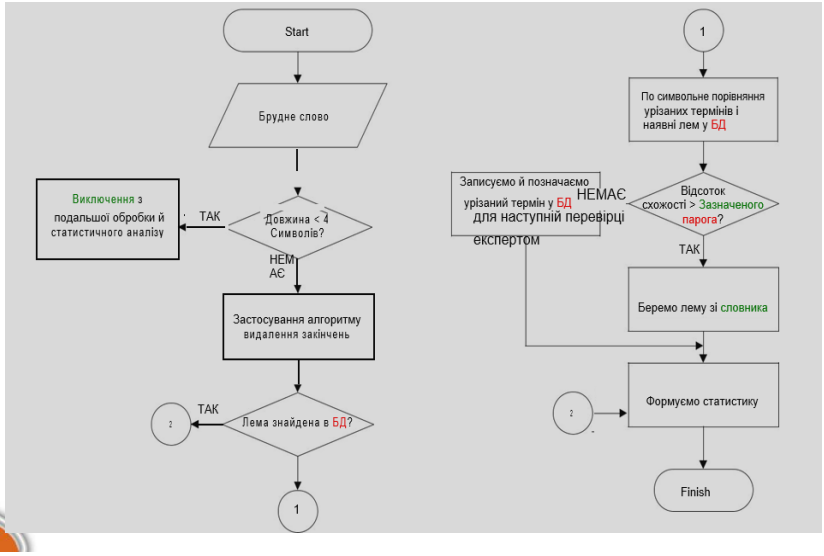
8

Процес лінгвістичної обробки запитів ІК та текстів ІР



9

Схема алгоритму лінгвістичної обробки термінів



10

Послідовність дій по перетворенню вихідного запиту ІК/тексту ІР у об’єкт, придатний для подальшої обробки алгоритмами кластерного аналізу

- виділення всіх термінів із запиту ІК/тексту ІР;
- стемінг і отримання урізаних термінів після видалення закінчень;
- перевірка урізаних термінів по словникові лем БД. Якщо лема знайдена, перехід до пункту д).
- збереження позначених урізаних термінів, для яких не визначена лема із БД. При необхідності лінгвістичний експерт може підтвердити або змінити позначені урізані терміни, перетворюючи їх у нові леми;
- формування статистики термінів і характеристичних векторів.

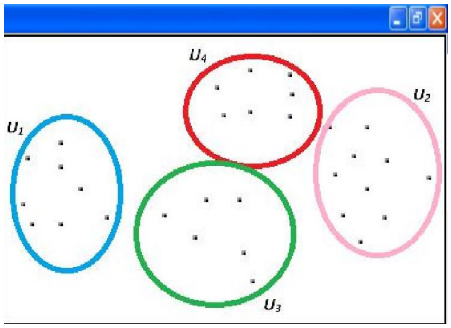
11

Визначення найближчого сусіда $U_{near}(t_{k+1})$ дозволить ініціалізувати місце положення нового об'єкта в новій кластерній структурі.

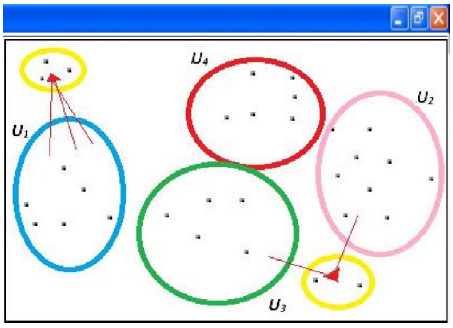
$$\rho(u_{nof(U(tk))+p}(t_{k+1}), u_{near}(t_{k+1})) = \min_{1 \le i \le nof\ (U(t_k))} \rho\left(u_{nof(U(t_k))+p}(t_{k+1}), u_i(t_{k+1})\right).$$

12

Ілюстрація розподілу об'єктів по кластерах у момент часу t_k



Ілюстрація розподілу об'єктів по кластерах у момент часу t_{k+2}



13

Розробка алгоритмів зміни в структурі кластерів

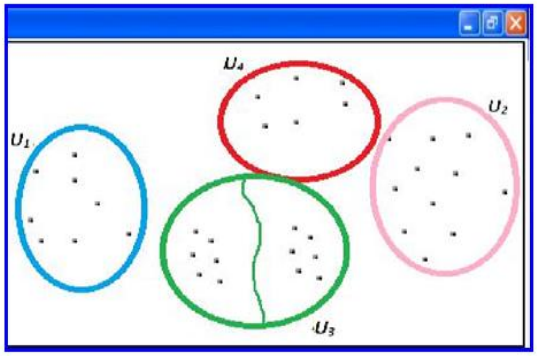
- Кількісним критерієм для оголошення двох кластерів кластерами, що зливаються, може бути їх подібності

$$I_{l,m} = \frac{\sum_{i=1}^{nof(U)} \min(b_{i,l}, b_{i,m})}{\sum_{i=1}^{nof(U)} b_{i,l}},$$

14

Алгоритм зміни в структурі кластерів

- Ілюстрація розщеплення кластера **U3** і формування всередині нього двох розділених згустків в t_{k+4}

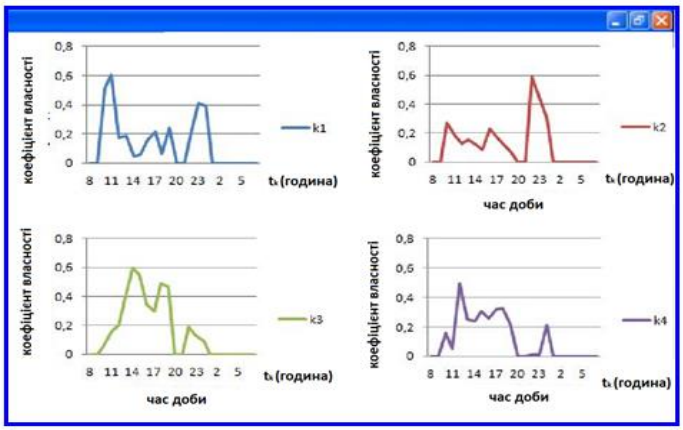


$$R = \frac{f_{\max} - f_{\min}}{f_{\max}} \geq g$$

15

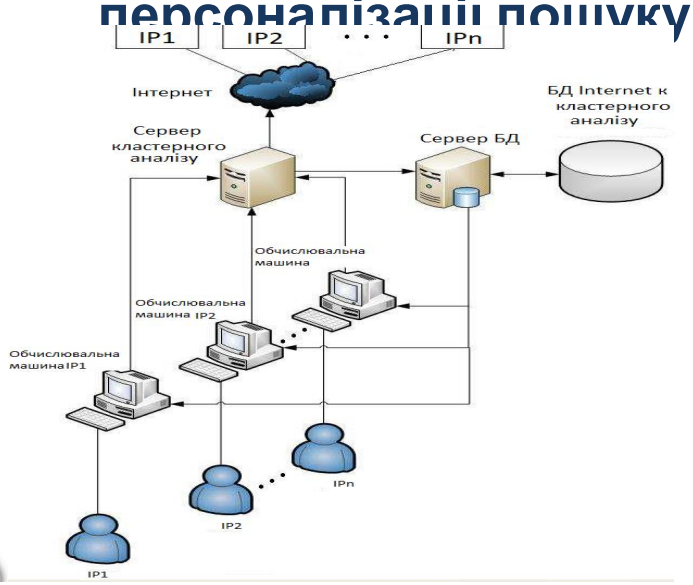
Перехід від динамічної до статичної кластеризації

- Графіки зміни коефіцієнта приналежності користувача до різних кластерів у різні моменти часу



16

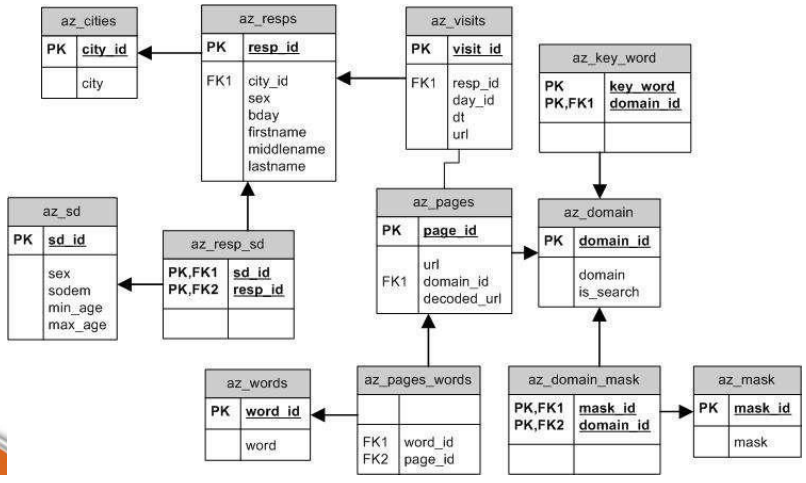
Розроблена структура корпоративної системи персоналізації пошуку



17

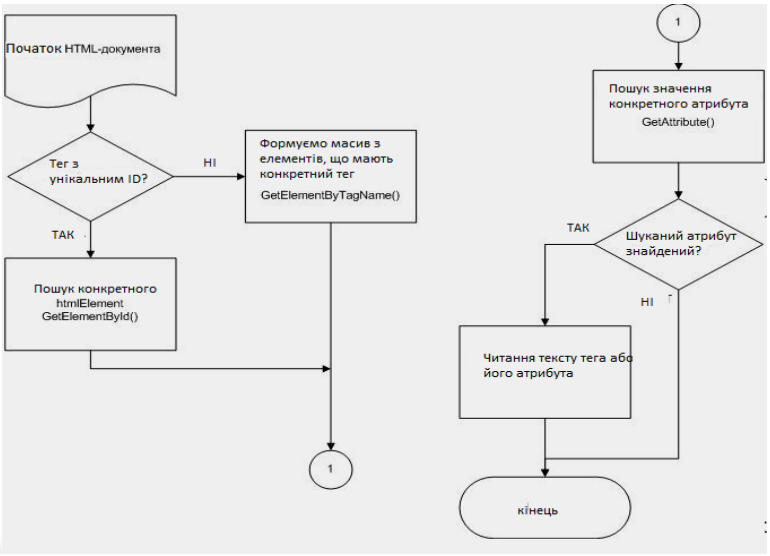
Додавання сутностей az_words і az_pages_words

- Повна структура БД для обробки заходів ІК



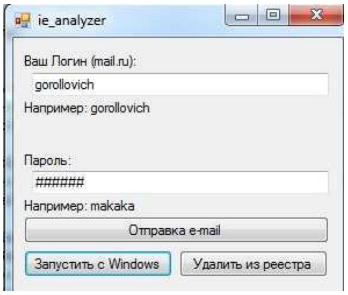
18

Схема алгоритму доступу до Dom-елементів



19

Графічний інтерфейс програми ie_analyzer



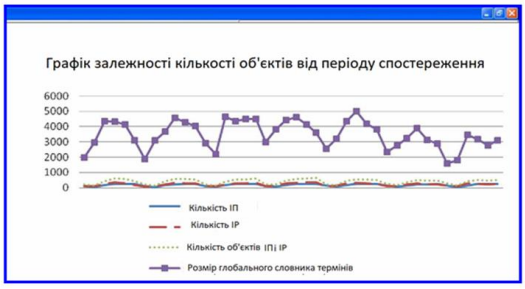
20

Алгоритм класифікації нових об'єктів

- У більшості своїх кроків повторює кроки алгоритму ініціалізації об'єктів і їх первісного розподілу по кластерах.
- Це свого роду одна чергова ітерація процесу кластерного аналізу.
- По завершенню цієї процедури отримуємо відсоток влучення в цільову групу, тобто відношення числа ІК, які насправді відвідали хоча б один з нових кластеризованих об'єктів до числа ІК, що беруть участь у кластеризації

21

Графіки залежності кількості об'єктів від періоду спостереження



22

Графіки залежності відсотків потрапляння в цільову групу і кластеризації від періоду спостереження при k = 2 і Δt = 4 години

23

Гістограма випадання

Випадання F вважається гарним показником для аналізу продуктивності пошукової системи

24

Аналіз дослідної експлуатації

- Отримані показники для пошукової системи в першу чергу пов'язані з тим, що дослідження проводилося винятково на перших 50 гіперпосиланнях (немає можливості кластеризувати усі 3000000 IP).
- Не дивлячись на це, у результаті проведених експериментів, можна вважати доведеним перевагу застосування кластерного аналізу для персоналізації пошуку.
- Слід зазначити, що значення зазначених показників будуть збільшуватися усе більше й більше, прагнучи до 1, при зростанні активності ІК.

Висновки

- Запропонована схема кластеризації IP зі зворотним зв'язком. Реалізація схеми дозволяє перетворювати динамічні IP у статичні IP і застосовувати до останніх стандартні алгоритми кластерного аналізу.

Пропонований підхід може бути використаний на рівні корпоративної мережі й охоплює практично всі етапи рішення цільового завдання:

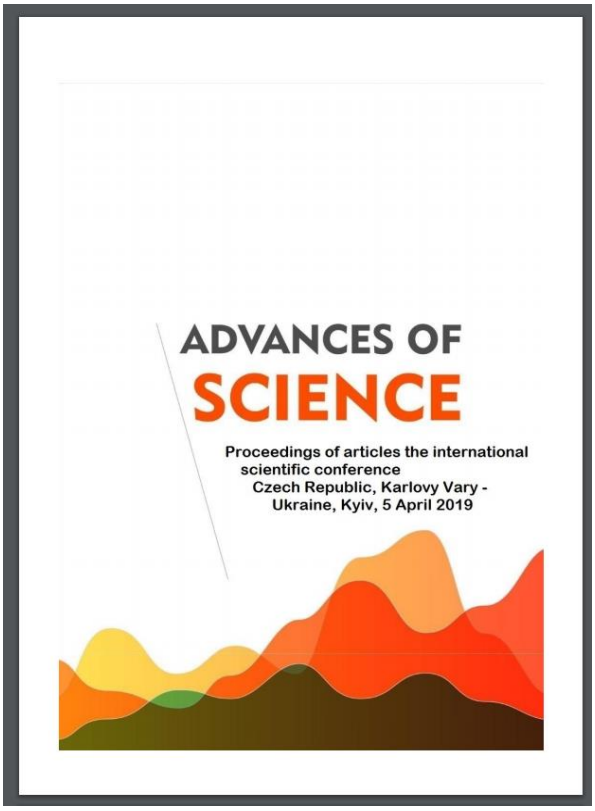
- первинний збір інформації, у межах заданого ковзного тимчасового вікна, про пошукову активність ІК і відвідуваних ними IP;
- багаторазове сканування Dom-моделі IP, застосування числових коефіцієнтів підсилення й Dom-фільтрації;
- структуризацію об'єктів за допомогою спеціалізованої БД;
- формування глобальних словників термінів і лем;
- формування характеристичних векторів ІК та IP, а також характеристичних векторів узагальнених Інтернет-об'єктів;
- розрахунок знаходження близькості між досліджуваними об'єктами;
- первісну ініціалізацію об'єктів;
- кластеризацію узагальнених об'єктів на основі алгоритму *k*-

Висновки (продовження)

- Використання результатів кластеризації для персоналізації пошуку – результати аналізу пошукової активності ІК у поточному інтервалі часу можуть бути застосовані для прогнозу його інформаційних потреб у наступних інтервалах часу.
- Розроблений набір програмних модулів для спостереження за активністю ІК і одержання текстового змісту IP з обліком їх Dom-моделей. Розроблені спеціальні збережені процедури, що виконують усі необхідні розрахунки – від формування словників термінів до кінцевого розподілу об'єктів по кластерах.
- Зазначені модулі й збережені процедури утворюють єдину програмну систему, яка, бувша встановленою на сервери локальної мережі, дозволить, наприклад, організувати на підприємстві корпоративну систему персоналізації пошуку

ДОДАТОК В

Апробація роботи



ADVANCES OF SCIENCE

Proceedings of articles the international scientific conference Czech Republic, Karlovy Vary – Ukraine, Kyiv, 5 April 2019

Czech Republic, Karlovy Vary – Ukraine, Kyiv, 2019

МЕТОД КЛАСТЕРІЗАЦІЇ РЕСУРСІВ ІНФОРМАЦІЙНИХ МЕРЕЖ

БАТКОВА Т.В.
студент, магістрант
Харківський національний університет радіоелектроніки
м. Харків, Україна

ШУБІН І.Ю.
канд. техн. наук, професор кафедри Програмної інженерії
Харківський національний університет радіоелектроніки
м. Харків, Україна

Величезна кількість Інтернет-ресурсів та інформації, що зберігається утримується в них, перетворило всесвітню павутину в грандіозне сховище погано організованих, неструктурованих даних. Пошук інформації в мережі Інтернет став долею людства. Будь-який інформаційний пошук (ІП) формує свою систему класифікації й відбору веб-ресурсів для задоволення потреб користувача в інформації. Користувач Інтернету має свій особистий психологічний портрет і відвідує конкретні, «улюблені» їм веб-сторінки. Якщо говорити про поведінку людини в мережі Інтернет, то можна виділити короткочасні (сесійні) дії ІП, які пов'язані з пошуком конкретної інформації протягом однієї або кілька пошукових сесій. Коли користувач знаходить релевантну інформацію, він прІПиняє свій пошук і навіть може вийти з мережі. Крім сесійних дій користувачів можна виділити їхня рутинна поведінка в мережі, наприклад, щоденний ранковий огляд новин про спорт або спілкування в соціальних мережах.

Великі пошукові системи користуються персональною інформацією та файлами *cookie* із браузерів для персоналізації результатів пошуку, – наприклад, маркетологи підбирають рекламу залежно від пошукової історії або залежно від статі й віку ІП. Вдаліше всього застосовується регіональний або

географічний таргетінг люди думають, що пошукова система дійсно порозумнішав і сказати, що це не так, не можна. Насправді пошукові системи добре працює з регіональними запитами при пошуку магазинів/товарів місцевого користування/споживання. Програмісти працюють над алгоритмами, що підвищують релевантність документів запитам за допомогою розрахунку ваг пошукових термінів, що дозволяє відбирати релевантні результати й переваги користувачів. У компанії Яндекс крім лінгвістичного аналізу контенту, індексу цитування, функції *DCG* (*Discounted cumulative gain*) [1], системи машинного навчання [2] і фільтрів негативних ознак у число таких методів входять і різноманітні процедури обліку й обробки первинної персональної інформації. Коли користувачі видають запити, приблизно в 20% випадків вони формулюють запити неоднозначно [1].

Декларуються різні цілі застосування методів кластерного аналізу до Інтернет-об'єктів, однак у переважній більшості випадків деталі цих методів і способів їх застосування не розголошуються. Так у роботах [3, 4] відзначається, що для кластеризації текстової інформації можуть використовуватися методи *TF* і *TIDF*, а також їх модифікації. Ці методи дійсно підходять для кластеризації текстів газет, підручників, наукових статей і інших інформаційних ресурсів зі статичним змістом. Наведені аргументи свідчать про необхідність подальшого пристосування Інтернету до потреб користувачів і, зокрема, за рахунок персоналізації Інтернет-пошуку. Підвищення рівня персоналізації пошуку, у свою чергу, може бути досягнуте за рахунок розробки перспективних методів класифікації ІП і ІР, заснованих на кластерному аналізі, впровадження цих методів в існуючі пошукові системи.

Проблема: відсутність ефективних методів і засобів, що забезпечують персонализацию пошуку інформації в Інтернеті. персоналізації пошуку жаркі дискусії йдуть уже майже 20 років – усі зацікавлені в тому, щоб результати пошуку в Інтернеті були як можна більш релевантними користувацьким запитам. Однак недостатня наукова опрацьованість проблеми, закритість більшості практично реалізованих рішень провідними компаніями

IENCE_05.04.19.pdf

постачальниками Інтернет-послуг обумовила необхідність дослідження теоретичних і практичних питань застосування методів кластерного аналізу для персоналізації пошуку. По темі кластерного аналізу існує велика кількість джерел. Вона охоплює загальні питання математичного опису об'єктів і алгоритми їх кластеризації. Кластеризація об'єктів зі статичними властивостями широко застосовується повсякденно в основному в аналітичній діяльності. Однак, методи кластеризації динамічних об'єктів, таких як ІР, недостатньо розроблені й, крім того, мало хто з дослідників розглядав ідею узагальненого представлення об'єктів різної природи, що мають подібні властивості.

Завданням роботи є застосування методів кластерного аналізу для класифікації ІП і ІР, для персоналізації інформаційного пошуку в Інтернет. Для досягнення поставленої мети потрібно розв'язати наступні основні завдання: проаналізувати існуючі методи кластерного аналізу ІП і ІР, показати їхні переваги в порівнянні з некластернимі методами та, запропонувавши адекватний математичний опис об'єктів дослідження, вибрати алгоритм кластеризації ІП і ІР, що дозволяє управляти результатом за допомогою вхідних параметрів.

Розроблено і застосовано оригінальний підхід, заснований на принципах узагальнення й одночасної кластерної обробки ІП і ІР.

Перехід від вербальних даних до числового представлення координат векторів відбувається за рахунок позиційного кодування термінів і підрахунку числа їх входжень у спостережуваний текст. Після векторизації інтернет-об'єктів проводиться розрахунок заходу близькості між ними й в остаточному підсумку формуються кластери з використанням одного із відомих алгоритмів.

Як результат натурних експериментів проведений вибір алгоритму кластеризации ІП і ІР, що забезпечує найкращі показники кластерної структури – їм виявився алгоритм *k-середніх*. Саме по собі завдання персоналізації пошуку досить старе і у комерційних цілях вже широко застосовуються некластерні методи класифікації ІП, що засновані на статичній інформації, також асоціативні методи класифікації ІР. Однак ці методи не враховують

інтереси ІП і якість класифікації ІР. Існуючі методи кластеризації текстів не беруть до уваги особливості сучасних ІР: не враховуються динамічні компоненти Dom-моделей ІР. Завдання пошуку оптимального підходу до кластеризації повинне враховувати, як поведінку ІП, так і динаміку ІР. Слід звернути увагу на той факт, що кластеризація ІП і ІР зараз проводиться роздільно.

Нехай X множина усіх спостережуваних об'єктів $x_i \in X$, $1 \le i \le nof(X)$, віднесених до одного із кластерів $X_l \subset X$, $1 \le l \le nof(K)$, де $K = \{X_1, \ldots, X_l, \ldots, X_{nof(K)}\}$ – множина усіх сформованих кластерів. У різні моменти часу t_k, $k = 0,1,2, \ldots$ проводяться спостереження за зміною стану кластерної структури залежно від характеристик об'єктів x_i, при цьому стан кожного *i-го* об'єкта в довільний момент часу t_k відображається характеристичним вектором $z_i(tk)$. Тут необхідно говорити про часову складову як додатковий параметр для всіх елементів вектора, що характеризує об'єкт. Якщо об'єкт дослідження при ієрархічної кластеризації представлений вектором $z_i = (z_{i,1}, \ldots, z_{i,j}, \ldots, z_{i,n})$, який не залежить від часу, то в динамічній системі кластеризації необхідно говорити про вектор $z_i(tk) = (z_{i,1}(tk), \ldots, z_{i,j}(tk), \ldots, z_{i,n}(tk))$, координати якого прив'язані до моментів часу t_k. У будь-який фіксований момент часу t_k (або інтервал часу Δt_k) можна виділити кілька кластерів, усередині яких об'єкти мають загальні характеристики. Зміна характеристик об'єкта $u_i \in U$ у момент часу t_k може привести до глобальних змін на рівні всієї кластерної структури й тим самим через період часу Δt_k буде необхідно провести нову кластеризацію всіх об'єктів з U. Якщо після витікання часу Δt_k кількість кластерів, їх зміст, розміри й положення їх центрів не змінюються, то мова може йти про так звану статичну кластерну структуру. Однак зовсім інакше буде в ситуаціях, коли із плином часу кластерна структура змінюється, коли об'єкти із часом починають мати деякі нові характеристики й утворюють групу об'єктів, функціонування яких перебуває на границі кластерів або навіть за його межами. У такому випадку кластерна структура перетерплює часові зміни й стає динамічною.

Слід подивитися на кластеризацію, як на завдання моніторингу сукупності інтернет-об'єктів з *n-мірним* характеристичним вектором числових ознак $z_i(tk)$, де індекс i відповідає номеру об'єкта. Вимір характеристик даних об'єктів здійснюється в дискретні, не обов'язково рівновіддалені моменти часу t_k. Через інтервал часу Δt_k проводиться перевірка стабільності кластерної структури й при необхідності (при наявності динамічних змін) її корекція. Наприклад, для кластеризації ІП, з метою виключення впливу занадто старих спостережень, моніторинг стану кластерів доцільно організувати за принципом часового вікна, тобто в $nof(Vu)$-мірному просторі при аналізі кластерної структури мають враховуватися тільки об'єкти, що зафіксовані в останньому часовому вікні Δt_k. Якщо говорити про часове вікно для обліку нових інтернет-об'єктів, то можна припустити наступні динамічні зміни в структурі кластерів: утворення нових кластерів, злиття кластерів, розщеплення або дроблення кластерів, зникнення кластерів, переміщення центрів кластерів.

Запропонований метод забезпечує зовсім новий підхід і дає нову математичну модель узагальнення ІП і ІР як єдиного об'єкта дослідження. Викладений метод може бути застосований не тільки для персоналізації пошуку в Інтернет, але й для вирішення широкого кола завдань, де є взаємодія людини з множиною подібних об'єктів, які необхідно класифікувати відповідно до їхніх властивостей.

Запропоновані методи можуть бути застосовані для дослідження й розробки пошукових систем загального й спеціального призначення, що мають високий рівень персоналізації пошуку. Прикладами таких систем може бути соціальна пошукова система, що працює на рівні вузькоспеціалізованих груп користувачів, або корпоративна система персоналізації пошуку, яка формує пошуковий результат залежно від пошукової спрямованості відділів підприємства – бухгалтерії, фінансового відділу, відділу маркетингу і т.ін.

Слід зазначити, що розроблені методи мають певні обмеження по застосуванню. Вочевидь, їх недоцільно застосовувати в умовах, коли одним комп'ютером (браузером) не користується більш одного користувача, тому що в

цьому випадку необхідне настроювання автоматичного очищення історії пошуку й *cookie* у самому браузері. Кожний ІП має свої пошукові інтереси й веде свій спосіб життя в кіберпросторі, тому різним ІП властиві різні інтереси й, як наслідок, вони можуть попадати в різні кластери.

Використана література

1. Барсегян А. А. Методы и модели анализа данных: OLAP и Data mining. / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. – СПб. : БХВ-Петербург. – 2004.
2. Ночевнов Д. Методы и средства сегментации web-сайтов // XVth International Conference "Knowledge-Dialogue-Solution" KDS-2. 2009 // Електронний ресурс // URL: http://www.foibg.com/ibs_isc/ibs-15/ibs-15-p13.pdf
3. Разделяй и властвуй: кластерные поисковики // Електронний ресурс // UPGRADE твой компьютерный еженедельник: сетевой журнал . URL: http://www.upweek.ru/razdelyaj-i-vlastvuj-klasternye-poiskoviki.html.
4. Библиотека работы с DOM HTML-документов для C# // Електронний ресурс // URL: http://htmlagilitypack.codeplex.com/
5. Bien J., Tibshirani R. Hierarchical Clustering With Prototypes via Minimax Linkage // Journal of the American Statistical Association. 2011 // Електронний ресурс // URL: http://faculty.bscb.cornell.edu/~bien/papers/jasa2011minimax.pdf

ДОДАТОК Г

Електронні матеріали