5. Висновки

Таким чином, створення та ведення геоінформаційної системи дозволяє значно збільшити якість обслуговування населення, організовувати ремонтні та профілактичні роботи, вести аналіз проведених та запланованих робіт.

6. Література

1. Бобровский С.А., Щербаков С.Г., Яковлев Е.И., Гарляускас А.И., Грачев В.В. Трубопроводный транспорт газа. - М.: «Наука», 2007. – с.3-495.
2. Серпинас Б.Б. Глобальные системы позиционирования. - М.: ИКФ «Каталог», 2002. – с.106.
3. Савиных В.П., Цветков В.Я. Геоинформационный анализ данных дистанционного зондирования. - М.: Картгеоцентр – Геоиздат, 2001. – с.228.
4. Стаскевич Н.Л. Справочное руководство по газоснабжению. - М.; Л.: Гостоптех издат, 2005. – с.217.
5. Вайсфельд В.А, Ексаев Л.Р. Принципиальные основы применения ГИС-технологий для городских инженерных коммуникаций // Инженерные коммуникации и геоинформационные системы: материалы первого учебно-практического семинара, «ГИС-Ассоциация», 14-17 октября 1997 г. – М.: 1997, - с.3-9.

В даній статті розглядається байєсівська класифікація текстових документів. Для більш детального аналізу та застосування на практиці розглядається новий підхід для кластеризації документів – ієрархічна кластеризація на основі частих наборів (FIHC), яка грунтується на ідеї частих наборів

Ключові слова: байєсівська класифікація, документ, текст, програма, об'єкт, клас, часті набори, кластеризація

В данной статье рассматривается байесовская классификация текстовых документов. Для более детального анализа и применения на практике рассматривается новый подход для кластеризации документов – иерархическая кластеризация на основе частых наборов (FIHC), основанная на идее частых наборов

Ключевые слова: байесовская классификация, документ, текст, программа, объект, класс, частые наборы, кластеризация

This article considers Bayesian classification of text documents. For a more detailed analysis and practical use of the new approach for document clustering - hierarchical clustering based on frequent sets (FIHC), which is based on the idea of frequent recruitment

Keywords: Bayesian classification, document, text, program, object, class, frequent sets, clustering

УДК 001.891:65.011.56

БАЙЄСІВСЬКА КЛАСИФІКАЦІЯ ТЕКСТОВИХ ДОКУМЕНТІВ. ІЄРАРХІЧНА КЛАСТЕРИЗАЦІЯ НА ОСНОВІ ЧАСТИХ НАБОРІВ

Я.С. Свириденко
Харківський національний університет радіоелектроніки
пр-т Леніна, 14, м. Харків, Україна
Контактний тел.: 096-951-87-18
E-mail: Kisssim@mail.ru

1. Вступ

У сучасний час різними сховищами знань накопичені величезні інформаційні масиви. Проблема полягає в складності орієнтування в цих масивах, у відсутності можливості отримати найбільш актуальну та повну інформацію по конкретній темі, що робить марною більшу частину накопичених ресурсів. Класифікація та кластеризація текстових документів являються одними з можливих варіантів вирішення проблеми використання інформаційних ресурсів.

Методи класифікації та кластеризації текстів застосовуються у фільтрації документів, розпізнаванні спаму, автоматичному анотуванні, складанні інтернет-каталогів, класифікації новин, розподілі реклами, персональних новинах. Також ці методи застосову-

ється у багатьох програмних системах для: фільтрації електронної пошти, контролювання новин, виборчого поширення інформації для споживачів, автоматичної індексації наукових статей, автоматичного пошуку категорій web-ресурсів, ідентифікації документів [1].

2. Постановка завдання

У даній статті розглядається байєсівська класифікація текстових документів, а також новий підхід для кластеризації документів – ієрархічна кластеризація на основі частих наборів (FIHC), який грунтується на ідеї частих наборів і запропонований R. Agrawal та R. Srikant [2]. Критерієм такої кластеризації є те, що в документі встановлюються деякі часті набори для кластера (теми), і різні кластери мають частку кількох частих наборів. Останній метод програмно реалізован, що дозволяє побачити його в дії на реальних прикладах.

3. Аналіз байєсівської класифікації

Байєсівська класифікація є широким класом алгоритмів класифікації, що заснована на принципі максимуму апостеріорної ймовірності. Для об'єкту, що класифікується, обчислюються функції правдоподібності кожного із класів, по яких обчислюються апостеріорні ймовірності класів. Об'єкт відноситься до того класу, для якого апостеріорна ймовірність максимальна. Цей алгоритм має мінімальну ймовірність помилок. Важливою особливістю байєсівської класифікації є те, що вона може бути побудована на основі вибірки із пропущеними значеннями. Також слід зазначити, що даний класифікатор виходить із припущення про те, що наявність або відсутність значень змінних носить абсолютно випадковий характер.

Завдання класифікації визначається в такий спосіб. Є безліч об'єктів X, не обов'язково кінцевих, і так само безліч $C=\{c_i\}$, де $i=1...N_c$, що складаються з N_c класів об'єктів. Кожен клас c_i представлений деяким описом F_i, що має деяку внутрішню структуру. Процедура класифікації f полягає у виконанні перетворень над об'єктами $x \in X$, після яких або робиться висновок про відповідність x одній із структур F_i, що означає віднесення x до класу c_i, або висновок про неможливість класифікації x. Елементами X є електронні версії текстових документів. Загальна модель текстового класифікатора може бути представлена основною алгебраїчною системою наступного виду:

$$R=<X,C,F,R_c,f>, \qquad (1)$$

де X – тексти, що підлягають класифікації, C – класи, F – опис класу, R_c – відношення на $C \times F$, f – операція рубрикування виду $X \to C$.

Крім сформульованого завдання класифікації, визначається завдання навчання рубрикатора, під яким мається на увазі часткове або повне формування C, F, R_c й f на основі деяких апріорних даних.

Згідно з виразом (1) текстові класифікатори можуть бути розділені залежно від способу подання описів класів (внутрішня структура елементів безлічі F) та від організації процедури класифікації f. У даній роботі буде розглядатися байєсівська класифікація, що відноситься до статистичних класифікаторів, на основі ймовірних методів. Загальною рисою для таких текстових класифікаторів є процедура f, в основі якої лежить формула Байєса для умовної ймовірності.

Аналізований текст x представляється у вигляді послідовності термінів $\{w_k\}$. Кожен клас c_i характеризується безумовною ймовірністю його вибору $P(c_i)$ у процесі класифікації деякого документа ($\sum P(c_i)=1$) і умовною ймовірністю $P(w|c_i)$ зустріти термін w у документі x за умови вибору класу c_i. Ці величини утворять елементи F_i безлічі F описів класів і будуть використані при розрахунку ймовірностей $P(x|c_i)$ того, що текст буде класифікований за умови вибору класу c_i. При розрахунку $P(x|c_i)$ враховується подання x у вигляді послідовності термінів w_k.

Підстановка цих величин у формулу Байєса дає ймовірність того, що буде обраний клас c_i, за умови, що документ x пройде успішну класифікацію (2). Процедура f зводиться до підрахунку $P(c_i|x)$ для всіх класів c_i і вибору того, для якого ця величина максимальна. Навчання зводиться до складання словника $\{w_n\}$ і визначення для кожного класу $P(c_i)$ і $P(w|c_i)$, де $w \in \{w_n\}$.

$$P(c_i|x)=\frac{P(c_i)*P(x|c_i)}{\sum P(c_i)*P(x|c_i)}. \qquad (2)$$

До числа байєсівських методів класифікації відносяться: наївний байєсівський класифікатор, лінійний дискримінант Фішера, квадратичний дискримінант, метод парзенівського вікна, метод потенційних функцій, логістична регресія, байєсівська мережа довіри.

4. Ієрархічна кластеризація документів на основі використання частих наборів

Даний метод працює в декілька кроків, включаючи видалення стоп-слів, і відбувається на наборі документів. Стоп-слова – це набір артиклів, таких як the, a, in, of і т.д. Кожен документ подається вектором частот інших пунктів документа. У методі FIHC в центрі уваги знаходиться кластер тому, що вимірюється узгодженість кластеру безпосередньо за допомогою частих наборів: документи на однакову тему містять більше загальних частих наборів, ніж документи під різними темами. Глобальний частий набір являє собою набір елементів, які будуть з'являтися в більш ніж мінімальній частині цілого набору документів. На це вказується мінімальна глобальна підтримка у відсотках від загальної кількості документів.

Даний метод будує кластери у два етапи: спочатку побудова початкового кластеру, а потім побудова початкового непересічного кластеру.

Побудова початкових кластерів. Для глобального частого набору будується початковий кластер, який містить всі документи, що й даний набір. Початкові кластери не перетинаються, оскільки один документ може містити кілька глобальних частих наборів. Основною властивістю початкових кластерів є те, що всі документи кластеру містять усі елементи глобального частого набору, який визначає кластер.

Побудова непересічних кластерів. Для документа визначається «кращий» початковий кластер і цей доку-

мент зберігається лише в найкращий початковий кластер. Визначається функція оцінки $Scope(C_i <- doc_j)$. Кластер C_i є «гарним» для документа doc_j, якщо є багато глобальних частих елементів в doc_j, що з'являються в «багатьох» документах в C_i. Оцінка заходів для покращення початкового кластера C_i для документа doc_j має такий вигляд:

$$Score(C_i \leftarrow doc_j) = \left[\sum_x n(x) * cluster\_support(x)\right] - \left[\sum_{x'} n(x') * global\_support(x')\right], \quad (3)$$

де x – глобальний частий елемент в doc_j і елемент частого кластеру в C_i; x' – глобальний частий елемент в doc_j, що не зустрічається в частому кластері C_i; $n(x)$ та $n(x')$ – зважені частоти x і x' у векторі ознак; $n(x)$ та $n(x')$ визначають $TF \times IDF$ на елементі x і x'.

Перший термін функції оцінки покращує кластер C_i, якщо глобальний частий елемент x у doc_j є в кластері C_i. Другий термін функції «карає» кластер C_i, я це глобальний частий елемент x' у doc_j, що не належить кластеру в C_i.

Розглянемо дванадцять документів, які наведено в табл. 1. Ці документи вибираються з набору документів [4] й їх назви вказують на їхні теми. Кожен документ заданий вектором ознак. У табл. 2 знаходяться всі глобальні часті k-набори з їхньою глобальною підтримкою (мінімальна глобальна підтримка = 35%). Початкові кластери наведені в табл. 3 (мінімальна підтримка кластера = 70%).

Таблиця 1

Перелік документів

Назва документу	Вектор ознак					
	(flow,	form,	layer,	patient,	result,	treatment)
cisi.1	(0	1	0	0	0	0)
cran.1	(1	1	1	0	0	0)
cran.2	(2	0	1	0	0	0)
cran.3	(2	1	2	0	3	0)
cran.4	(2	0	3	0	0	0)
cran.5	(1	0	2	0	0	0)
med.1	(0	0	0	8	1	2)
med.2	(0	1	0	4	3	1)
med.3	(0	0	0	3	0	2)
med.4	(0	0	0	6	3	3)
med.5	(0	1	0	4	0	0)
med.6	(0	0	0	9	1	1)

Таблиця 2

Глобальні часті набори

Глобальний частий набір	Глобальна підтримка
{flow}	42%
{form}	42%
{layer}	42%
{patient}	50%
{result}	42%
{treatment}	42%
{flow, layer }	42%
{patient, treatment }	42%

Таблиця 3

Початкові кластери

Кластер (ярлик)	Документи у кластері	Часті елементи кластеру та підтримка кластера (CS)
C(flow)	cran.1, cran.2, cran.3, cran.4, cran.5	{flow, CS=100%}, {layer, CS=100%}
C(form)	cisi 1, cran.1, cran.3, med.2, med.5	{form, CS=100%}
C(layer)	cran.1, cran.2, cran.3, cran.4, cran.5	{layer, CS=100%}, {flow, CS=100%}
C(patient)	med.1, med.2, med.3, med.4, med.5, med.6	{patient, CS=100%}, {treatment, CS=83%}
C(result)	cran.3, med.1, med.2, med.4, med.6	{result, CS=100%}, {patient, CS=80%}, {treatment, CS=80%}
C(treatment)	med.1, med.2, med.3, med.4, med.6	{treatment, CS=100%}, {patient, CS=100%}, {result, CS=80%}
C(flow, layer)	cran.1, cran.2, cran.3, cran.4, cran.5	{flow, CS=100%}, {layer, CS=100%}
C(patient, treatment)	med.1, med.2, med.3, med.4, med.6	{patient, CS=100%}, {treatment, CS=100%}, {result, CS=80%}

Для того, щоб знайти найбільш підходящий кластер для документа med.6, треба розрахувати його оцінки щодо початкового кластера, що містить цей документ:

$$Score(C(patient) \leftarrow med.6) = 9*1+1*0.83-1*0.42 = 9.41$$

$$Score(C(treatment) \leftarrow med.6) = 10.8$$

$$Score(C(result) \leftarrow med.6) = 9$$

$$Score(C(patient, treatment) \leftarrow med.6) = 10.8$$

Для пояснення розрахунку, використаю оцінку $Score(C(patient) \leftarrow med.6)$. Глобальні часті елементи в med.6 – це «patient», «result», і «treatment». Їхні частоти вектора ознак 9, 1 і 1 відповідно. «Patient « і « treatment « є частими кластерами в кластері $C(patient)$, отже, ці два елементи з'являються в частині функції покращення та їхні частоти перемножаються на відповідну підтримку кластера 1 і 0,83 відповідно. «Result» не зустрічається у кластері $C(patient)$, тому він з'являється в частині покарання і його частота множитися на його глобальну підтримку, що становить 0,42.

Обидва кластера $C(treatment)$ і $C(patient, treatment)$ отримали одну й ту саму високу оцінку. Документ med.6 призначається до $C(patient, treatment)$, який має більше число елементів у кластерному ярлиці, тобто кластер з більш конкретною темою. Після повторення цих обчислень для всіх документів, отримую непересічні кластери (табл. 4).

Таблиця 4

Непересічні кластери

Кластер (ярлик)	Документи у кластері	Часті елементи кластеру та підтримка кластера (CS)
C(flow)	cran.1, cran.2, cran.3, cran.4, cran.5	{flow, CS=100%}, {layer, CS=100%}
C(form)	cisi. 1	{form, CS=100%}
C(layer)		none
C(patient)	med.5	{patient, CS=100%}, {treatment, CS=83%}
C(result)		none
C(treatment)		{treatment, CS=100%}, {patient, CS=100%}, {result, CS=80%}
C(flow, layer)		none
C(patient, treatment)	med.1, med.2, med.3, med.4, med.6	{patient, CS=100%}, {treatment, CS=100%}, {result, CS=80%}

Оскільки деякі документи видаляються з початкових кластерів, потрібно перерахувати часті елементи кластеру для кожної групи. При повторному обчисленні частих елементів кластеру C_i включаються всі документи з усіх «нащадків» C_i. Кластер є нащадком C_i, якщо його кластерний ярлик являється розширенням кластерного ярлика C_i. Ідея полягає в тому, що нащадки будуть мати підрозділи тем батьків, тому їх треба включати. У табл. 4 третій стовпець відображає оновлення частих елементів кластеру в непересічних кластерах. Підтримка кластера елемента «treatment» у кластері C(patient) становить 83%, оскільки п'ять із шести документів містять цей елемент.

5. Побудова дерева кластера

Набір кластерів, побудованих на попередньому етапі, можна розглядати як набір тем і підтем в документі. Дерево кластера (тем) побудоване на основі подібностей між кластерами. У випадку, коли дерево містить занадто багато кластерів, застосовуються два методи скорочення для ефективного скорочення й звуження дерева шляхом об'єднання аналогічних кластерів.

Дерево кластера будується знизу вгору, вибираючи батьків на рівні k – 1 для кластера на рівні k. Для k-кластера C_i на рівні k, спочатку треба виявити всіх потенційних батьків, які є (k – 1)-кластерами і мають кластерний ярлик, який є підмножиною кластерного ярлика в C_i. Існує принаймні k таких потенційних батьків. Наступним кроком є вибір «найкращого» серед потенційних батьків. Спочатку треба об'єднати всі документи в піддереві C_i в єдиний документ $doc(C_i)$, який поступово вводитися в побудову дерева знизу вгору. Потім обчислюється оцінка $doc(C_i)$ щодо потенційного батька. Потенційний батько з найбільшою кількістю очок стане батьком C_i. Всі листя кластерів, які не містять жодного документа, можуть бути видалені.

Розглянемо кластери в табл. 4. Починаю будувати дерево з двох кластерів (наприклад, кластери з двох наборів у кластерному ярлиці). Кластер C(flow,layer) видаляють, оскільки він є порожнім вузлом. Далі треба вибрати батьків для C(patient,treatment). Потенційними батьками є C(patient) і C(treatment). C(treatment) отримує більш високу оцінку і стає батьком C(patient,treatment). На рис. 1 показані результати дерева кластеру.

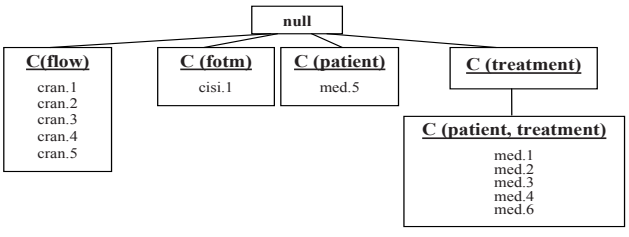

Рис. 1. Дерево кластера

Скорочення дерева. Скорочення дерева полягає в об'єднанні аналогічних кластерів з метою отримання звичайної ієрархії тем для перегляду й підвищення точності кластеризації. Спочатку треба визначити міжкластерну подібність, яка є ключовим поняттям для об'єднання кластерів.

Для вимірювання міжкластерної подібності між двома кластерами C_a та C_b потрібно виміряти схожість C_b на C_a і схожість C_a на C_b. Ідея полягає в розгляді одного кластеру як концептуального документа (шляхом об'єднання всіх документів у кластері) і вимірюванні його оцінки щодо інших кластерів за допомогою функції оцінки. Оцінка повинна бути нормована щодо видалення зміни розміру документа. Схожість C_j на C_i визначається наступним чином:

$$Sim(C_i \leftarrow C_j) = \frac{Score(C_i \leftarrow doc(C_j))}{\sum_x n(x) + \sum_{x'} n(x')} + 1, \qquad (4)$$

де C_j і C_i є двома кластерами; $doc(C_j)$ виступає для об'єднання всіх документів у піддереві C_j в єдиний документ; x – глобальний частий елемент в $doc(C_j)$ і частий кластер в C_i; x' – глобальний частий елемент в $doc(C_j)$ і не є частим кластером в C_i; n(x) – зважена частота x у векторі ознак $doc(C_j)$; n(x') – зважена частота x' у векторі ознак $doc(C_j)$. Після нормалізації оцінки за $\sum_x n(x) + \sum_{x'} n(x')$ нормовані значення знаходяться в діапазоні [–1,1]. У результаті, діапазон функції Sim становить [0,2].

Міжкластерну схожість між C_a та C_b визначається як середнє геометричне двох нормованих оцінок $Sim(C_a \leftarrow C_b)$ і $Sim(C_b \leftarrow C_a)$:

$$Inter\_Sim(C_a \leftrightarrow C_b) = \left[Sim(C_a \leftarrow C_b) * Sim(C_b \leftarrow C_a)\right]^{\frac{1}{2}}, \qquad (5)$$

де C_a та C_b є двома кластерами, які включають своїх кластерних потомків.

Спосіб «Скорочення дитини». Мета цього методу – ефективно скоротити дерево за допомогою заміни кластерів дитини на кластери її батьків. Критерій скорочення заснований на міжкластерній схожості батька й дитини. Дитина скорочується тільки тоді, коли вона схожа на свого батька.

Процедура сканування дерева проходити знизу вгору. Для не листового вузла розраховується

Inter_Sim між вузлом і кожним з його дітей, і скорочується, якщо дитина кластеру Inter_Sim вище 1. Якщо кластер скоротився, то його діти стали дітьми своїх предків. Скорочення дитини застосовується тільки до рівня 2.

Для визначення кластеру C(patient,treatment) його потрібно скоротити. Міжкластерна схожість C(treatment) і C(patient,treatment) розраховується так:

$$\mathrm{Sim}(C(\mathrm{patient}) \leftarrow C(\mathrm{patient},\mathrm{treatment})) =$$
$$= (30*1+9*1+8*0.8-1*0.42)/48+1=1.94$$

$$\mathrm{Sim}(C(\mathrm{patient},\mathrm{treatment}) \leftarrow C(\mathrm{treatment})) =$$
$$= (30*1+9*1+8*0.8-1*0.42)/48+1=1.94$$

$$\mathrm{Inter\_Sim}(C(\mathrm{patient}) \leftrightarrow C(\mathrm{patient},\mathrm{treatment})) =$$
$$= (1.94*1.94)^{\frac{1}{2}} = 1.94$$

Для розрахунку $\mathrm{Sim}(C(\mathrm{patient}) \leftarrow C(\mathrm{patient},\mathrm{treatment}))$, треба об'єднати всі документи в кластері C(patient,treatment) шляхом додавання їх векторів ознак. Вектора ознак має вигляд (0, 1, 0, 30, 8, 9). Потім обчислюється оцінка цьої комбінації документів C(treatment) і вона нормалізується за сумою частот, яка становить 48.

$\mathrm{Sim}(C(\mathrm{patient},\mathrm{treatment}) \leftarrow C(\mathrm{treatment}))$ обчислюється з використанням того самого методу. З міжкластерною схожістю вище 1 кластер C(patient,treatment) скорочується (рис. 2).

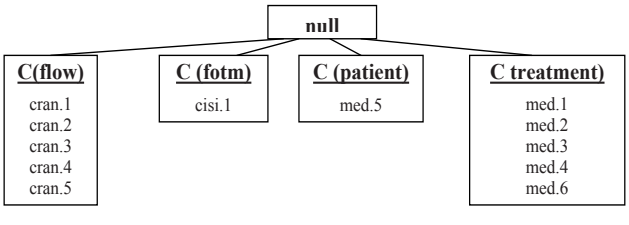

Рис. 2. Дерево кластера після методу «Скорочення дитини»

Метод «Споріднене злиття». Так як метод «Скорочення дитини» застосовується тільки до рівня 2 і нижче, то часто залишається багато кластерів на рівні 1. «Споріднене злиття» об'єднує аналогічні кластери на рівні 1. Кожного разу розраховується Inter_Sim для кожної пари кластерів на рівні 1 і об'єднуються ті пари кластерів, які мають найвище значення Inter_Sim. За рахунок такого об'єднання діти з двох кластерів стають дітьми об'єднаного кластера. Ця процедура повторюється для кластерів на рівні 1 до тих пір, доки не буде вичерпано вказану користувачем кількості кластерів. Якщо користувач не вказує кількість кластерів, то алгоритм завершиться, коли всі пари кластерів на рівні 1 матимуть значення Inter_Sim нижче або дорівнюватимуть 1. Попарне порівняння гарантує те, що об'єднуються тільки аналогічні кластери.

Метод «Споріднене злиття» розраховує Inter_Sim для кожної пари кластерів на рівні 1 (табл. 5).

Таблиця 5

Розрахунок міжкластерної схожості

Пари кластерів (C_i, C_j)	$\mathrm{Sim}(C_j \leftarrow C_i)$	$\mathrm{Sim}(C_i \leftarrow C_j)$	$\mathrm{Inter\_Sim}(C_i \leftrightarrow C_j)$
C(flow) & C(form)	0.71	0.58	0.64
C(flow) & C(patient)	0.58	0.54	0.56
C(flow) & C(treatment)	0.75	0.53	0.63
C(form) & C(patient)	0.58	0.58	0.58
C(form) & C(treatment)	0	0.56	0
C(patient) & C(treatment)	1.72	1.70	1.71

Якщо користувач не вказав потрібну кількість кластерів, FIHC припинив би своє виконання і повернув би дерево, яке зображене на рис. 2. Наприклад, задане число кластерів 2. Алгоритм скоротить один кластер на рівні 1 на основі кластерної схожості між кластерами C(flow), C(form), C(patient) і C(treatment).Так як C(patient) і C(treatment) пара з високим значенням Inter_Sim, тим менше кластерів C(patient) об'єднається з більшим кластером C(treatment). На рис. 3 зображене результуюче дерево.

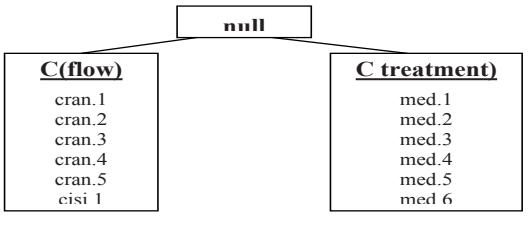

Рис. 3. Дерево кластера після методів «Скорочення дитини» та «Споріднене злиття»

6. Опис програмної системи

Програмна система, що реалізує метод FIHC, будує ієрархію кластеризованих документів з набору некластеризованих документів.

Спочатку програма зчитує файли з вказаної директорії, потім проводить кластеризацію документів по описаному в попередньому розділі алгоритмі. В результаті роботи даної програми користувач отримує XML-файл, в якому представлена ієрархія кластеризованих документів. Для генерації XML-файлу програма використовує Microsoft XML Parser.

7. Висновки

В даній статті було розглянуто байєсівську класифікацію, а також підхід для кластерізації документів – ієрархічна кластерізація на основі частих наборів (FIHC), який заснований на ідеї частих наборів і запропонованій R. Agrawal та R. Srikant. Останій підхід був розглянутий на конкретному прикладі.

Література

1. Автоматична класифікація текстів [Електронний ресурс] / Юрій Лифшиц. – Режим доступу: http://logic.pdmi.ras.ru/~yura/internet/06ia.pdf – 25.11.2006 р. – Загол. з екрану.

2. Agrawal, R. Fast algorithm for mining association rules [Текст] – L.: Proc. 20th Int. Conf. Very Large Data Bases, 1994. – 487 p.
3. Classic [Електронний ресурс] / K. Readman. – Режим доступу: ftp://ftp.cs.cornell.edu/pub/smart/ – 7.11.2002 р. – Загол. з екрану.
8. Вапник, В. Н. Теорія розпізнавання образів [Текст] / В. Н Вапник, А. Я Червоненкис. – М.: Наука, 1974. – 598 с.
9. Прикладна статистика: класифікація й зниження розмірності [Текст] / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. – М.: Фінанси й статистика, 1989. – 457 с.

Запропонований інтегральний скалярний показник якості передачі мовних сигналів в пакетних мережах з втратами, заснований на пофонемному спектральному аналізі. На основі введеного показника досліджена залежність якості передачі мови від параметрів мережі

Ключові слова: пофонемний короткочасний спектральний аналіз мовних сигналів, мережевий емулятор

Предложен интегральный скалярный показатель качества передачи речевых сигналов в пакетных сетях с потерями, основанный на пофонемном спектральном анализе. На основе введенного показателя исследована зависимость качества передачи речи от параметров сети

Ключевые слова: пофонемный кратковременный спектральный анализ речевых сигналов, сетевой эмулятор

A proposed integrated scalar index of quality of voice signals over packet networks with losses, based on spectral analysis of speech signals

Through the introduction parameters the dependence of voice quality from the network properties has been researched

Keywords: spectral analysis of speech signals, the network emulator

УДК 621.391

ОЦЕНКА ПОКАЗАТЕЛЕЙ ПОФОНЕМНОГО СПЕКТРАЛЬНОГО АНАЛИЗА РЕЧЕВЫХ ФРАГМЕНТОВ В IP – СЕТЯХ

С.М. Бобрицкий
Заведующий лабораторией
Харьковский институт судебных экспертиз
им. засл. проф. Н.С. Бокариуса
ул. Золочевская, 8а, г. Харьков, Украина, 61177
Контактный тел.: (057) 777-67-33
E-mail: Bobriski@hniise.gov.ua

М.Ю. Ощепков
Доцент*
Контактный тел.: (057) 702-13-20
E-mail: Oshchepkov@kture.kharkov.ua

В.Е. Саваневич
Доктор технических наук, доцент, профессор*
*Кафедра телекоммуникационных систем
Харьковский национальный университет
радиоэлектроники
пр. Ленина, 14, г. Харьков, Украина, 61166
Контактный тел.: (057) 702-55-92
E-mail: Domsv1@rambler.ru

1. Введение

Одним из основных составляющих трафика современных IP –сетей является передача речи в режиме коммутации пакетов [1, 2, 3, 4]. При этом проблемным при организации передачи речи в реальном масштабе времени, является удовлетворение повышающихся со временем требований к качеству.