УДК 342.5+35.078.3

Олег Анатолійович Хатян

ТЕХНОЛОГІЧНІ ЗАСАДИ ПОБУДОВИ МОДЕЛІ АНАЛІЗУ ПОТОКІВ ІНФОРМАЦІЙНИХ ПОВІДОМЛЕНЬ

Інформаційні потоки становлять основу інформаційного суспільства і значною мірою відображають тенденції розвитку та проблематику як загальнолюдського, так і національного характеру. Як свідчить світова практика, їхній аналіз дає змогу оперативно, в режимі реального часу, відслідковувати потреби суспільства й оптимально направляти зусилля та матеріальні ресурси на їх задоволення. На сьогодні, в умовах безмежного глобального інформаційного простору, спостерігається постійне зростання обсягу інформаційних ресурсів (мають-ся на увазі як глобальні ресурси, так і персональні інформаційні сховища організацій) і, як наслідок потужність потоків. За цих обставин робота інформаційно-аналітичних підрозділів вимагає застосування технічних засобів оброблення інформаційних потоків. Такі засоби покликана надати галузь інформаційних технологій. Але внаслідок безсистемності розвитку саме інформаційних ресурсів у цій галузі виникло кілька проблемних аспектів. Серед них:

- велика кількість надлишкової не відповідно класифікованої за тематикою інформації (якщо така класифікація існує взагалі);
- пошукові засоби пропонують обмежені за своїми можливостями мови запитів для відбору інформації з масивів;
- англомовність більшості доступних програмних засобів для роботи з текстовою інформацією;
- майже повна закритість технологій та структур побудови інформаційних і пошукових систем, що не дає можливості їх вдосконалення та адаптації під вирішення спеціалізованих завдань інформаційно-аналітичної роботи.

Викладене зосереджує увагу й робить актуальною проблему саме технологічних аспектів побудови спеціальних прикладних засобів автоматизації аналітичної діяльності.

Аналіз останніх досліджень і публікацій. Над вирішенням зазначених проблем працюють численні колективи вчених і фахівців усього світу. Так, консорціум W3C розробляє концепцію Семантичного Web [1], яка реалізується технологією Web-2 (http://www.web2con.com/) — "Web другого покоління". Зокрема Web-2 включає реалізацію семантичного Web, багаторівневу підтримку метаданих, нові підходи до дизайну і відповідного інструментарію, технологію глибинного аналізу текстів (Text Mining) [2], що базуються на статистичному аналізі з використанням ЕОМ [3; 4; 5; 6], а також ідеологію Web-сервісів, базуючись при цьому на інформаційних ресурсах, накопичених у WWW першого покоління.

Відомі підходи до окремих питань зазначеної проблеми і вітчизняних авторів. Так, А.В. Анісімов [7] розглядав проблеми аналізу літературних творів та текстових повідомлень із використанням методів комп'ютерної лінгвістики. Д.В. Ланде [8] запропонував використовувати методи фрактального та кластерного аналізу для вивчення інформаційних потоків.

Проте весь комплекс проблем до сьогодні не вирішений.

Отже, *метою статті* є формування технологічної основи побудови спеціальних прикладних засобів автоматизації аналізу інформаційних потоків. В досягненні мети допоможе вирішення таких завдань: побудова формальної моделі процесу класифікації інформаційних повідомлень та формулювання загальних принципів класифікації повідомлень.

Основна частина. Задачу побудови моделі процесу класифікації інформаційних повідомлень можна віднести до класу задач KDT (Knowledge Discovering in Text — пошук знань у тексті). Реалізація програмних пакетів за цією технологією перебуває в початковому стані [2], а наявні реалізації (наприклад, безкоштовні програми WordNet — семантична мережа, що містить більше 100 тис. слів англійської мови і зв'язує їх відносинами "синонім", "антонім" і т.д.) в

основному англомовні з інтелектуальними можливостями, що ґрунтується на розширеному пошуку частини рядка в тексті. Іншим продуктом є пакет ManageGigabyte (MG), в якому наявні функції компресії/декомпресії текстової інформації при занесенні в сховище та побудові індексних таблиць для реалізації контекстного пошуку. Цей продукт призначений для організації персонального інформаційного сховища із засобами пошуку інформації за ключовими словами і представлення тематичних класифікацій текстової інформації. Крім того, останній пакет є умовно безкоштовним і доступний із текстами програми.

Продовження проекту ManageGigabyte став програмний пакет GreenStone (GS), де реалізовані всі базові функції MG та додана підтримка Unicode.

Відомі підходи до вирішення задачі побудови структури представлення знань. Взагалі ефективне вирішення задач аналізу неструктурованих даних значною мірою залежить від вибору форми представлення знань. Загальні методи і підходи в цьому напрямку наведені в статті "Розвідка і дані" [2]. Серед них:

1. Дерева рішень (Decision Trees — DT). Це програма на примітивній мові програмування з мінімальною кількістю конструкцій, подібно тих, що використовуються в експертних систем (ЕС). Але в ЕС правила застосовувалися для представлення знань людей, формалізованих шляхом DT, а в KDD — подібні DT використовуються для представлення набору правил, що створені самонавчальною програмою.
2. Асоціативні правила (Association Rules — AR). В той час, як DT містять знання про відповідність між значеннями деяких атрибутів даних і критерійним показником, AR — використовується для формалізації знань про взаємозв'язок між різними атрибутами.
3. Правила вірогідності (Probablistic Rules — PR) — задають оцінки вірогідності ухвалення кожного правила в DT (на відміну від DT за, якої проводиться двійкова оцінка — "Так", "Ні").
4. Кластер — форма представлення знань, яка дає змогу з деякою мірою впевненості казати про схожість групи об'єктів. Знання про кластер є визначальним і зазвичай використовується для побудови "тонких" знань, що співвідноситься з реальним світом.
5. Нейронні мережі — певна чорна скринька, яка має вхідний та вихідний інтерфейс і властивість навчання. Навчання відбувається шляхом подання зовнішніх наборів сигналів, що характеризують об'єкти: "вплив" — на вхідний інтерфейс та "реакція" — на вихідний. Вважається, що після навчання нейронна мережа здатна вирішувати завдання класифікації об'єктів. (Розглянуто тільки один тип нейронної мережі).
6. SOM (карти самоорганізації Коханена) — карти самоорганізації Self-Organized Map — метод формалізації знань.
7. Поліноміальная форма представлення знань — Метод групового обліку аргументів (МГОА) використовувався в прогнозуванні (програма AbTech).
8. Геометричні методи ідентифікації — базуються на теорії ідентифікації систем за їх топологічними характеристиками.

Алгоритми розрахунку оціночних показників моделей інформаційних потоків

Методи пошуку знань у тексті базуються на алгоритмах математичного оброблення статистичних даних відносно лінгвістичних конструкцій, що містяться в текстових повідомленнях. Розгляду цих методів присвячені монографії з прикладної статистики. Серед них методи шкалювання, автоматичної класифікації об'єктів у багатовимірному просторі ознак, багатопараметричні дискримінантний та факторний методи аналізу статистичних даних, методу перевірки статистичних гіпотез та інші [3—6].

Прикладні системи та недоліки, що їм притаманні

Зазвичай для вирішення проблем впорядкування та пошуку в інформаційних масивах використовується програмне забезпечення, до якого відносяться інформаційно-пошукові системи (ІПС).

Аналіз, з урахуванням особливостей застосування в рамках аналітичної діяльності, в умовах конкурентного середовища, відомих на сьогодні ІПС дає змогу визначити низку істотних недоліків, які не дозволяють використовувати їх безпосередньо, оскільки:

1. ІПС є досить складним комплексом програм і потребує високої кваліфікації персоналу, що його обслуговує.
2. Тексти програмного забезпечення та використані алгоритми є інтелектуальною власністю розробника і повністю недоступні для адаптації.
3. Структури даних не доступні для модифікації.
4. Не виключаються "недокументовані можливості", що може призводити до зовнішнього несанкціонованого втручання та витоку інформації.
5. Як правило, існує англомовність набору правил опису мовних конструкцій.

6. Висока вартість пакетів аналізу даних у великих інформаційних масивах.
7. Майже всі пропоновані системи, що орієнтовані на організацію електронних систем діловодства та аналізу даних, вимагають участі фахівців організацій постачальників програмного забезпечення у створенні засобів зберігання інформації та доступу до технологічного циклу її оброблення.

Досвід відомих комерційних організацій, наприклад консалтингових, показує, що значна частина штату організації є висококваліфікованими фахівцями з розробки програмного забезпечення, орієнтованого на розроблення та постійне вдосконалення власного програмного продукту. Такі продукти досить спеціалізовані і не підлягають розповсюдженню.

Ця робота покликана розглянути один із можливих підходів до оброблення інформації, що подається в текстовому вигляді (текстові документи, інформаційні та повідомлення електронної пошти і т. ін.).

Іншими словами, предмет дослідження — інформаційні масиви (кажучи взагалі — потоки), які містять неструктуровану або погано формалізовану інформацію. Інформаційні потоки розглядаються з погляду на те, що запит довільного елементу може здійснюватись у будь-який момент часу, а наявність потоку покладає підвищені вимоги до швидкості його оброблення.

Загальний принцип побудови моделі

Взявши за мету формування технологічної основи автоматизації аналізу інформаційних потоків, ми розглядаємо ключовий прикладний аспект моделі, а саме — встановлення семантичного статусу текстового повідомлення, що означає визначення його розташування в семантичному полі, на відміну від однозначного віднесення повідомлення до тієї або іншої тематики.

Таке семантичне поле визначається побудовою простору семантичних ознак оцінювання повідомлення. Кількість таких ознак зумовлює вимірність простору семантичних ознак. Для кожної ознаки задається оціночний градієнт у вигляді шкали оцінки відповідності тексту певній ознаці. Водночас, на кожній шкалі будується множина проміжків оціночних значень, де кожний із них відповідає певній тематиці (рис. 1). За такої побудови семантичного поля опису тематики відповідає вектор, довжина якого дорівнює вимірності простору семантичних ознак, та кожний елемент якого є проміжок оціночних значень за певною ознакою. Оцінене повідомлення також становить собою вектор значень за кожною ознакою. Тоді, процес визначення належності повідомлення до тематики полягає у вирішенні питання: "наскільки визначені для повідомлення показники збігаються з проміжками, визначеними тематикою за кожною ознакою?". У цьому полягає принцип побудови моделі.

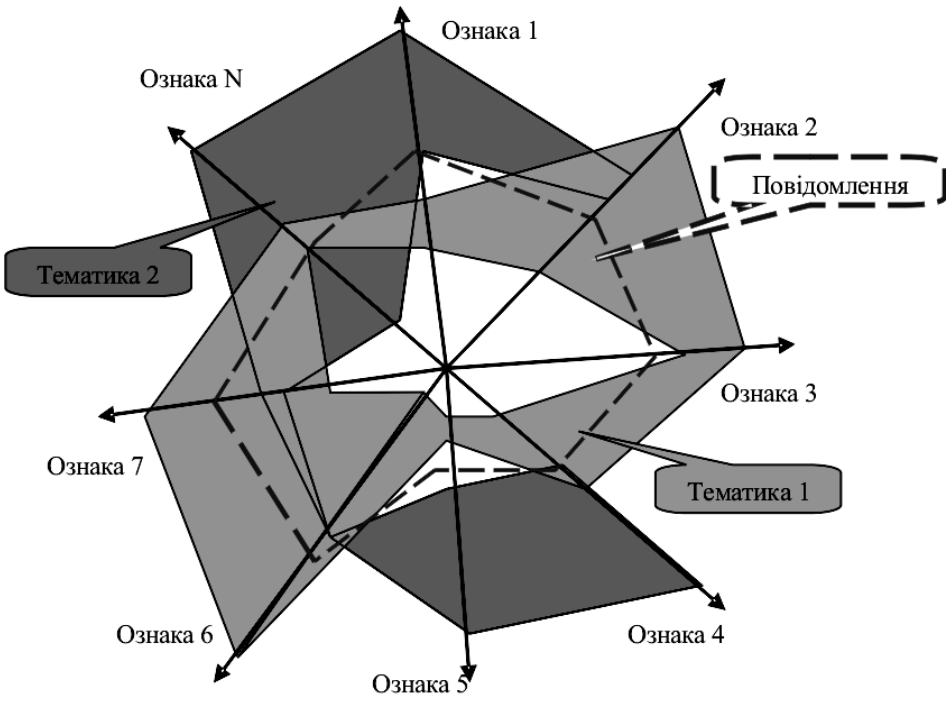

Рис. 1. Схематичне зображення простору семантичних ознак та визначення тематик, до яких належить повідомлення

Модель і основні поняття

Вважаємо, що інформаційний потік — це впорядкована за часом послідовність інформаційних повідомлень.

Інформаційне повідомлення — це набір інформації, представлений у текстовому вигляді (в подальшому — текст) як ланцюжок символів, кожний з яких має однозначний код у таблиці кодування. Беруться до уваги такі таблиці кодування win-1251 ASCII, 866 кодова сторінка з набором латинських, арабських, символів псевдографіки та цифр.

Стосовно тексту, його структурних та семантичних елементів розрізняємо такі визначення:

- слово — ланцюжок символів, що розділені символом із множини {“.”, “,”, “ ”, “;”, “:”, “...”, “?”, “!”, “<табуляція>”, “<кінець рядка>”};
- цифрове слово — ланцюжок символів-цифр разом із цифрами римської символіки {“I”, “V”, “X”, “L”, “C”, “D”, “M”}, що може починатись знаками {“+”, “–”}, та може містити символ десятинної крапки {“.”, “,”}; структурно це слово, тобто для виділення з тексту використовується та сама множина розподільних символів;
- символічне слово — ланцюжок алфавітно-цифрових символів, що може починатись знаками з множини {“№”, “€”, “$”, “&”, “*”}, містить один або кілька символів з множини {“.”, “,”, “+”, “–”, “±”, “=“, “ч”, “Ч”, “?”, “=“, “÷”, “\”, “/”, “{”, “}”, “[”, “]”, “(”, “)”, “^”} та може закінчуватись знаками з множини {“€”, “$”, “%”, “#”, “@”, “~”, “*”, “©”, “®”, “™”};
- речення — ланцюжок символів, що відділений символом із множини {“.”, “...”, “?”, “!”}; структурно складається зі слів;
- абзац — ланцюжок символів, що відділений одним або кількома символами початку параграфа; структурно складається з речень;
- текст — змістовно завершений документ, позначений атрибутивними характеристиками; становить кінцевий ланцюжок символів, який структурно складається з абзаців;
- атрибутивні характеристики тексту — обов’язкові: назва або тематична належність; необов’язкові: джерело, дата/час публікації, автор; як правило, структурно складається зі слів;
- об’єкт — слово або кілька слів, які є власною назвою, особи, явища, предмета;
- дія — слово або кілька слів, що семантично означає прояв деякої енергії, діяльності, а також сама сила, діяльність, функціонування;
- процес — сукупність послідовних дій стосовно об’єкта; відноситься до семантичного наповнення тексту.

Вважаємо, що кожний текст за визначенням є носієм думки, звідки природно випливає зв’язок символічного представлення об’єкта (лінгвістична складова) з його семантичним наповненням, як об’єкта реального світу. Наступне загальне зауваження є наслідком попереднього й полягає в міркуванні стосовно самого підходу до класифікації інформаційного повідомлення. Для ефективного автоматизованого аналізу інформаційних потоків достатньо створити таку модель інформаційного потоку, яка б була здатною за лінгвостатистичними показниками (отримуються з текстів та відповідають зазначеним дефініціям) характеризувати належність повідомлення до певної групи (збігається з першим етапом побудови інтерпретатора експертної системи). Необхідні ж умови задовольняє виконання другого етапу.

Отже, автор не має на меті створення генератора текстів. Ідея полягає в класифікації (в загальному випадку — розпізнаванні) об’єктів та процесів що містяться в текстовому повідомленні. Для вирішення цього завдання пропонується модель процесу класифікації. Ця модель полягає в такому:

- маємо потік вхідних повідомлень, визначених нами як тексти;
- кожне повідомлення підлягає впорядкуванню згідно з деякими правилами;
- результат впорядкування зберігається в інформаційному сховищі у вигляді текстів та пов’язаних із ними структур об’єктів;
- аналітики прикладної сфери формують у довільній формі запити стосовно об’єктів, їх зв’язків тощо;
- запит ототожнюється з текстовим повідомленням, а отже множина таких запитів становить собою потік;
- запити інтерпретуються згідно з тим же набором правил, що й тексти;
- результат оброблення запиту подається у формальному вигляді, тобто впорядкованій множині об’єктів із можливою деталізацією зв’язків та відповідним набором текстів.

Почнімо розгляд із понять тексту, об’єкта, їх характеристик, дій над об’єктами за текстом повідомлення, які можуть бути автоматично ідентифіковані. Узагальнену множину текстів та елементів, що їх становлять, назвемо системою. Будемо розглядати різні рівні взаємодій. У процесі опису моделі намагатимемося за можливості вказувати на математичний апарат, який дасть змогу адекватно формально описувати, аналізувати, класифікувати та моделювати поведінку об’єктів нашої системи.

Побудова формальної моделі

Формуємо мову моделі як певий словник первинних фактів (для моделі це значення набору вхідних параметрів, повний набір яких кількісно обмежений), тобто словники лексичних одиниць типу: об'єкт, дія, тематика. Тоді, значенням (або змістом) тексту або його частини назвемо відношення його до деякого класу належності (спільності), якому відповідає певна множина первинних фактів із загального словника. Множина класів належності збігається з тематичним рубрикатором прикладної сфери. Такі класи формуються за відповідними алгоритмами автоматично на базі тестових інформаційних одиниць — слів та їх зв'язків у тексті.

У моделі наявні:

1. Множина текстів.
2. Множина об'єктів, що містяться та описуються в тексті.
3. Множина властивостей об'єктів відповідно до їх опису в тексті.
4. Множина можливих дій стосовно об'єктів за текстом.
5. Множина відношень об'єктів, що виводяться на базі первинних фактів, яким відповідають множини, описані пунктами 2, 3, 4, та деякої множини аксіом.
6. Множина комплексних відображень (спотворень суті) об'єкта, що з'ясовуються в результаті додаткового аналізу.

Вважаємо кожну множину елементів моделі злічимою — це дасть нам можливість ранжувати й посилатися на них. Вважаючи множину значень канторовим, а множину елементів текстів і самих текстів — злічимою, гіпотетично можна встановити деяку міру відповідності конкретного тексту певному класу.

Зважаючи на концептуальні уявлення й застосовуючи введену термінологію, можна зробити висновок, що вся множина значень розбита на дві групи. До першої групи відносяться — чітко визначені тексти, що піддаються формальному опису відповідно до прийнятої системи параметрів, та обмежені зонами в багатовимірному просторі ознак, яким відповідає набір первинних фактів. У термінології моделювання процесів взаємовідношень об'єктів вказані зони відповідають множині стаціонарних станів, у яких можуть перебувати пов'язані об'єкти, та піддаються формальному опису. До того ж зв'язки таких об'єктів чітко виражені та характеризуються значною потужністю.

Наведений приклад повідомлення (Додаток 1) має досить чітко виражену множину об'єктів, дій і співвідношення до тематичних рубрик і, як наслідок досить просто підлягає класифікації.

До текстів другої групи (Додаток 2) відносяться такі, що важко підлягають формальному опису, і для їх аналізу найбільш придатними є напівавтоматичні або повністю ручні методи класифікації. Крім того, попередній аналіз дає змогу припустити, що частина текстів першої групи значно перевищує другу (якщо йдеться про інформаційні потоки, де кожний текст покликаний нести певне конкретне значення та орієнтований на комунікативну складову).

Взагалі кажучи, можна переформулювати поставлене завдання для побудови інтерпретатора експертної системи як деякого транслятора з урахуванням наведених понять таким чином.

Загальне завдання побудови інтерпретатора експертної системи поділяється на два етапи. Інтерпретуємо в наведених нами визначеннях тексту як послідовності первинних фактів.

Етап перший. Необхідно розробити правила, що дають можливість визначити, чи є деяка вхідна послідовність первинних фактів правильною фразою сформованої в моделі мови, тобто чи має така послідовність адекватний сенс у прикладній сфері (чи визначена семантична шкала при формуванні семантичного поля). Якщо це так, тоді необхідно визначити синтаксичну структуру послідовності в термінах внутрішньої мови опису значень.

Інакше необхідно зазначити, на якому етапі аналізу вхідної послідовності відсутня логічна інформація, що міститься в експертній системі у вигляді правил.

Етап другий. Полягає в побудові, на мові прикладної сфери (мові визначення тематик, що групують множину значень текстів, зрозумілій аналітикам), використовуючи деяку кількість формальних правил, висновку про належність тексту конкретній тематиці, заданій у запиті.

Отже, маємо дві складові частини моделі (в широкому значенні це дві повноцінні моделі з високим ступенем взаємної інтеграції).

Внутрішня — ґрунтується на принципах класифікації текстів за іменами і характеристиками об'єктів, описаних текстом. Класифікація проводиться за рівнями на перших етапах, вертикально на завершальних. Довільність у напрямі проведення класифікації забезпечує нормованість вагових коефіцієнтів для визначених нами елементів тексту.

Зовнішня — на принципах побудови експертної системи, що має власну машину побудови експертного висновку про належність тексту тій або іншій тематиці, набір експертних (сформованих аналітиком) правил відбору текстів на підставі повної або часткової вказівки первинних чинників

формально або неформально заданих (первинними чинниками є імена й характеристики об'єктів, одержаних у результаті аналізу інформаційного потоку).

Як прототип зовнішньої складової частини моделі можна розглядати процес роботи аналітика з текстом. Завдання аналітика, наприклад, виявляти тексти, що належать до одного об'єкта, і на основі узагальненої інформації формувати опис видів діяльності об'єкта. Формально, до властивостей системи на базі описаних принципів відносяться:

1. Часткове виявлення близьких за значенням коротких текстових повідомлень.
2. Встановлення зв'язків об'єктів згідно з тематичним рубрикатором (тематичні зрізи) на великому масиві текстової інформації.
3. Реалізація контекстного пошуку текстів за окремими словами та їх послідовностями.

У сенсі практичної реалізації моделі фактично завдання зводиться до:

1. Формалізованого представлення тексту, його окремих елементів та атрибутивних характеристик.
2. Моделювання поведінки оператора-аналітика в процесі віднесення тексту до того або іншого тематичного розділу, на основі формалізованої інформації про об'єкти та їх зв'язки за змістом.

У закінченому вигляді система, що створюється на основі вказаної моделі, є засобом:

1. Визначення правил, що описують знання (принципів конструювання текстів з елементів) експерта про об'єкти та їх відношення.
2. Розкладання текстів на складові та побудови контекстних індексів швидкого доступу до текстів.
3. Автоматичної класифікації текстів (завдання деякого чисельного значення або ранжирування текстів у деякому багатовимірному просторі встановлених чинників).
4. Представлення текстів у зручній структурі для зберігання і швидкого доступу з урахуванням знань про класи належності.
5. Підвищення ефективності аналізу потужних інформаційних потоків.

Висновки і перспективи подальших досліджень

Зважаючи на викладене вище, можна дійти висновку, що розглянута формальна модель та загальні принципи процесу класифікації інформаційних повідомлень, на базі оцінювання лінгвостатистичних показників, дають можливість сформувати технологічну основу побудови спеціальних прикладних засобів автоматизації аналізу інформаційних потоків.

Водночас, вказана технологія не дає змоги значною мірою адекватно класифікувати тексти довільної складності, з урахуванням великої кількості повідомлень в інформаційних потоках Інтернету. Цікавими в цьому напрямку є дослідження інформаційного потоку з погляду так званого фрактального уявлення про структуру об'єктів навколишнього середовища [8] та апарат і методологія, що надається відносно новим науково-філософським підходом — синергетикою [9].

Література

1. Berners-Lee T. The Semantic Web, Scientific American, May 2001 [Електронний ресурс] / Tim Berners-Lee, James Hendler, Ora Lassila Режим доступу: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21. **2. Розвідка** і дані // Компьютер. обозрение. — 2000. — № 12. — С. 36. **3. Афифи А.** Статистический анализ. Подход с использованием ЭВМ / А. Афифи, С. Эйзен. — М. : Мир, 1982. — 488 с. **4. Енюков И. С.** Методы, алгоритмы, программы многомерного статистического анализа (пакет ППСА). — (Матеем. обеспечение прикладной статистики) / И. С. Енюков. — М. : Финансы и статистика, 1986. — 232 с. **5. Факторный,** дискриминантный и кластерный анализ [Пер.: Jae-On Kim, Charles W. Mueller. Factor Analysis: Statistical Methods and Practical Issues. Sage Publication, Inc., 1987. William R. Klecka.Discriminant Analysis. Sage Publication, Inc., 1980. Mark S. Aldenderfer, Roger K. Blashfield. Cluster Analysis. Sage Publication, Inc., 1984]. — М. : Финансы и статистика, 1987. — 607 с. **6. Статистические** методы для ЭВМ / [под ред. К. Эйслейна, Э. Рэлстона, Г. С. Уолфа] / перевод с англ. — М. : Наука, 1986. — 460 с. **7. Анисимов А. В.** Компьютерная лингвистика для всех: Мифы. Алгоритмы. Язык / А. В. Анисимов. — К. : Наук. думка, 1991. — 208 с. **8. Ландэ Д. В.** Фракталы и кластеры в информационном пространстве. [Електронний ресурс] / Д. В. Ландэ Режим доступу: http://www.it2b.ru/it2b2.view2. page116.html. **9. Хакен Г.** Самоорганизующееся общество / Г. Хакен / переклад з нім. Е. Н. Князевої. Режим доступу: http://spkurdyumov.narod.ru/Haken51.htm.

Додаток 1

Приклад повідомлення з чітко вираженою множиною об'єктів, дій і співвідношенням до тематичних рубрик:

01.10.2008 22:53 Президент затвердив Директиви на переговори з РФ

Київ. 1 жовтня. УНІАН. Президент України Віктор ЮЩЕНКО своїм Указом затвердив Директиви на переговори під час робочого візиту прем'єр-міністра України Юлії ТИМОШЕНКО до Російської Федерації 2 жовтня 2008 року.

Як повідомили УНІАН у прес-службі Президента України, Директиви містять позицію української сторони щодо окремих питань економічного співробітництва України з РФ, зняття бар'єрів у взаємній торгівлі, умов поставок природного газу споживачам і його транзиту територією України та співробітництва в інших галузях.

Визначені об'єкти:

ПРЕЗИДЕНТ УКРАЇНИ, ВІКТОР ЮЩЕНКО
ПРЕМ'ЄР-МІНІСТР УКРАЇНИ, ЮЛІЯ ТИМОШЕНКО
УКРАЇНА
РОСІЙСЬКА ФЕДЕРАЦІЯ, РФ
ДИРЕКТИВА
Дії:
ЗАТВЕРДИВ
ВІЗИТ
ПЕРЕГОВОРИ
ЗНЯТТЯ БАР'ЄРІВ
ЕКОНОМІЧНЕ СПІВРОБІТНИЦТВО
ТОРГІВЛЯ
ПОСТАВКИ ПРИРОДНОГО ГАЗУ
ТРАНЗИТ ГАЗУ

Тематики, до яких відноситься повідомлення:

РОСІЯ
ПАЛИВНО-ЕНЕРГЕТИЧНА ГАЛУЗЬ — УКРАЇНА
ПРЕЗИДЕНТ УКРАЇНИ
КАБІНЕТ МІНІСТРІВ УКРАЇНИ
ПРОТИСТОЯННЯ ОРГАНІВ ДЕРЖАВНОЇ ВЛАДИ
ОГЛЯД ЗМІ З ГАЗОВИХ ПИТАНЬ

Додаток 2

Приклад повідомлення наукового характеру, що важко підлягають класифікації:

01.10.2008 23:08 Ученые назвали плюсы отцовского воспитания

stream: 'РосБизнесКонсалтинг'.

Ни для кого не секрет, что для любого ребенка важно воспитание обоих родителей. Однако недавно британские ученые доказали, что дети, в чьем развитии активную роль играет отец, вырастают более умными и успешными, сообщают британские СМИ. В течение 50 лет ученые из Центра поведения и эволюции при университете Ньюкасла следили за жизнью 17 тыс. младенцев, рожденных в одну и ту же неделю, анализируя влияние активного отцовского воспитания. Когда участники исследования достигли зрелого возраста, ученые провели с ними подробное интервью, в котором определили их социальную мобильность, успешность, наличие собственной семьи, а также то, насколько они являются лучшими родителями, чем их собственные.

Только в 2004 году были опрошены 5600 человек, достигших возраста 46 лет. Результаты опроса подтвердили полученные на протяжении всего исследования данные. Как выяснилось, дети, чьи отцы играли активную роль в их взращивании, совершая с ними частные прогулки и читая вместе книжки, выросли более успешными, нежели те, чье воспитание было возложено на плечи матерей.

Несмотря на то, что обычно отцы более склонны нянчиться с сыновьями, чем с дочками, их воспитание оказалось плодотворным на детей обоих полов. При этом ученые обнаружили, что рожденные от более старших мужчин дети более склонны страдать аутизмом. Результаты исследования, опубликованные в журнале "British Journal of Psychiatry", показали, что подобный риск повышается в три раза, если отцу ребенка во время зачатия больше 25 лет. Из всего количества детей 100 человек оказались отсталыми в развитии. Все они были рождены от мужчин солидного возраста.

01 октября 2008 г.

РИА "РосБизнесКонсалтинг" 2008.10.01 21:10

В статье рассматриваются технологические основы построения формальной модели анализа информационных сообщений. Приводится обзор алгоритмов, структур представления данных и недостатков прикладных систем, которые применяются для решения задач анализа сообщений. Указываются возможные перспективы дальнейшего совершенствования модели на основе новейших концептуальных подходов.

Ключевые слова: информационный поток, модель, классификация, анализ данных, анализ текстов, лингвостатистический анализ.

Technological bases of construction of formal model for analysis of information messages are examined in this article. A review of algorithms, structures of presented data and disadvantages of application-oriented systems which are used for solving the tasks of message analysis is given. Possible prospects of further improvement of the model on the basis of the newest conceptual approaches are specified.

Key words: information flow, model, classification, data analysis, text analysis, statistic-linguistic analysis.