5. Кращим засобом перевірки правильності моделі є її кількаразове тестування. Обираючи прикладний пакет для ІС, необхідно в першу чергу визначитися з інструментарієм бізнес-моделювання. В ідеалі цей інструментарій повинен бути вбудованим в програмний пакет, щоб бізнес-модель залишалася частиною ІС.

6. Для підтримання бізнес-моделі в актуальному стані (на всіх стадіях проекту впровадження ІС і після його завершення) необхідно, щоб усі задачі організаційного планування на підприємстві розглядалися в межах наявної бізнес-моделі.

1. Катренко А.В., Бобало І.Ю. Основні тенденції розвитку методів проектування інформаційних систем // Вісн. Держ. ун-ту "Львівська політехніка". – 1997. – № 315. – С. 50–71. 2. Катренко А.В. Стандарти у проектуванні та експлуатації інформаційних систем // Вісн. Держ. ун-ту "Львівська політехніка". – 2000. – № 406. – С. 135–155. 3. Implementing BAAN IV. Yves Perreault and Tom Vlasic. – 1998. 4. Business Process Oriented Implementation of Standard Software. Mathias Kirchmer. – 1998. 5. Darnton G., Darnton M. Business Process Analysis. – London: Tompson Business Press, 1997.

УДК 681.3

С.А. Катренко
Національний університет "Львівська політехніка",
кафедра "Прикладна лінгвістика"

ЗАСТОСУВАННЯ МЕТОДІВ НАВЧАННЯ ДЛЯ РОЗПІЗНАВАННЯ СПАМУ

This paper is concerned with problems of email handling and presents a memory-based approach of spam mail filtering. Although a lot of messages an average user receives are kind of unsolicated mail and there is a tendency of growth of these e-mails, the problem of mails filtering is still to be solved. Existing filters are not good enough to block all spam messages. Since many messages are in languages different than English we have chosen German as a language to work on. An overview of existing spam filtering approaches is given as well.

Розглянуто проблему опрацювання електронних повідомлень та запропоновано підхід машинного навчання для фільтрації небажаних повідомлень (спаму). Попри те, що багато з повідомлень, які отримує користувач, належать до категорії спаму та помітна тенденція зростання кількості таких повідомлень, ця проблема все ще не розв'язана. Існуючі фільтри недостатні для блокування усіх небажаних електронних повідомлень. Оскільки багато повідомлень написані різними мовами (і необов'язково англійською), обрано німецьку мову. Також подано огляд існуючих підходів фільтрації спаму.

ВСТУП

Кількість електронної пошти, не очікуваної користувачем, невпинно зростає. Водночас все більше повідомлень написані іншою мовою, аніж англійська. За оцінкою

margetagent.com, німецькомовна аудиторія отримує понад 500 млн. спам-повідомлень щотижня. Зокрема, кожен інтернет-користувач – два повідомлення щодня, а кожен четвертий – більше, ніж 20. Проте теми цих повідомлень не відрізняються від англомовного спаму, а саме, 72 % – еротика, 60,3 % – онлайн-торгівля, 71,2 % – флірт, пошук партнера, 44,4 % – позики. Найчастіше блокувати спам намагаються за допомогою фільтрів, котрі складаються з правил на зразок "якщо – то" і не завжди функціонують ефективно. Це спонукало до пошуку нових або застосування вже відомих методів.

ПОПЕРЕДНІ ДОСЛІДЖЕННЯ

Розглядаючи попередні дослідження, слід зауважити, що одні з них були обмежені лише до фільтрування спаму, позаяк інші класифікували усі електронні повідомлення.

Отже, методи та підходи, подані нижче, можна поділити на 2 групи: загальна класифікація електронної пошти та фільтрування спаму. Окрема група, а точніше, лише одне з досліджень, грунтувалося на абсолютно іншому підході – кластеризації електронних повідомлень [6] – і тому розглянуте окремо.

Перевагами застосування класифікаційних методів є використання різного типу інформації: текстової, що належить до різних частин повідомлення – тіла, теми, атачменту; числової інформації – скажімо, кількість отримувачів або булевих значень – наявність/відсутність атачментів.

Класифікація електронної пошти

Багато досліджень базуються на класифікації пошти згідно з заданими категоріями, що в загальному випадку відповідає папкам у поштовому клієнті. Серед підходів, які були застосовані для такого типу класифікації, виділяють класифікацію за допомогою машинного навчання та видобування інформації (machine learning and IR approaches).

До них належать:

1. MailCat (Segal, Kephart)

Автори використовують TF-IDF підхід, який полягає у обчисленні коефіцієнта TF-IDF як добутку частотності слів у кожній з категорій (tf) та відношення загальної кількості категорій до кількості категорій, в яких ці слова появляються. Під час класифікації нового повідомлення рівень помилок становив від 20 до 40 % .

2. Порівняння RIPPER та TF-IDF (Cohen)

У цьому підході була використана навчальна програма RIPPER, за допомогою якої продукувалися правила з використанням ключових слів повідомлень. Якщо слово, що входить у певне правило, з'являється у повідомленні, яке слід класифікувати, застосовується відповідне правило. Проте після порівняння цієї системи та TF–IDF зауважуємо, що результати не надто відрізняються (точність 87–94 % у першому випадку та 85–94 % для останнього)

3. SVM (support vector machines, Brutlag, Meek)

Brutlag та Meek також порівнювали ефективність використання TF-IDF, але цього разу, з SVM. Застосування SVM продемонструвало кращі результати для щільних категорій, а TF–IDF – для "розкиданих", "розріджених" (sparse data). Точність – для SVM 70–90 %, TF–IDF 67–95 %.

4. Re:Agent (Boone)

Опрацювання повідомлень під час застосування цього підходу здійснюється у два кроки: спершу формуються характеристики (features) або шляхом використання TF–IDF,

або шляхом явного відбору ключових слів користувачем. Далі ці характеристики застосовуються на етапі навчання у нейронних мережах або як дані для алгоритму визначення "найближчого сусіда" (nearest neighbour approach). Для класифікації використовувалося лише 2 категорії (work, other), з врахуванням цього результат є достатньо високим – 98 %.

Фільтрування та класифікація спаму

До підходів, які застосовувалися для фільтрування спаму, належать метод Баєса (Bayesian approach) та, як і у випадку класифікації поштових повідомлень, машинне навчання. Вони демонструють достатньо високу точність при визначенні спаму.

Більшість систем, побудованих за таким зразком, використовують попереднє опрацювання даних, а саме, ***токенізацію*** та ***стемінг*** (**tokenization and stemming**). На етапі токенізації визначаються токени, себто елементи, які будуть використовуватися під час навчання. Ними зазвичай є слова, тобто токенізація передбачає усунення пунктуації і екстрагування слів. Оскільки одне і те ж слово може використовуватися у різних формах (різних відмінках, числі тощо), необхідно здійснити стемінг, тобто виділити корінь/основу слова, або представити слова лише у інфінітиві (для дієслів), називному відмінкові однини (для іменників, займенників) тощо. Майже в усіх дослідженнях використовувалося вже розроблене програмне забезпечення, що здійснює стемінг.

Далі бажано видалити слова, що з'являються дуже часто (залежно від мови, прийменники, артиклі тощо).

Серед систем, які створені для фільтрації спаму і працюють за схожими схемами, можна виділити наступні:

1. SpamCop (Pantel, Lin) [7].

Ця система базується на застосуванні Naive Bayes-підходу й використовує стемінг та список слів (токенів), які необхідно видалити. До таких слів належать токени, які з'являються у всіх повідомленнях більш, аніж 4 рази, або ті, що розподілені однаково поміж спам-повідомленнями та приватною поштою. Автори цієї системи порівняли її ефективність з системою, розробленою Когеном (п. 2 у попередньому підрозділі), й продемонстрували, що SpamCop функціонує краще, а саме, з точністю до 94 %.

2. Підхід Баєса також використовували Sahami, Dumais, Heckerman та Horvitz.

Проте, на відміну від попередніх, ці автори визначали features двома шляхами – автоматично та кодування шляхом відбору певних токенів. Точність результатів лише з автоматично відібраними токенами становила 97,1 % для спаму та 87,7 % для приватних повідомлень. Повнота при цьому становила 94,3 % та 93,4 % відповідно.

3. Метод Баєса та навчання, що базується на аналогії [1]

Порівняльний аналіз ефективності застосування методу Баєса та підходу, що отримав назву memory-based learning approach був здійснений Androutsopoulos-ом. Цей автор класифікував повідомлення у дві категорії – "лінгвістичний спам" та "повідомлення з галузі лінгвістики". Тобто, дані були обрані так, аби вони належали до однієї і тієї ж теми (у цьому випадку, до лінгвістики). З них 481 повідомлення належало до категорії спам, 2412 повідомлень – до лінгвістичного корпусу. Вибір схожих за темою повідомлень ускладнює класифікацію, але водночас надає можливість реальніше оцінити ефективність обраного алгоритму та підходу. В межах цього аналізу автор застосовував різні параметри; для методу Баєса було отримано 98 % точності та 78 % повноти, у випадку застосування нав-

чання, яке базується на попередньо представлених зразках, 95,92 % та 85,27 % відповідно (для спаму). Точність результатів у випадку використання шаблонів Outlook дорівнювала 53,01 %, повнота при цьому становила 87,93 %.

Отже, як продемонстрували попередні дослідження, найкращі результати були досягнуті з застосуванням методу Баєса та навчання, що грунтується на зразках (для прикладу, точність результатів з застосуванням методу Баєса сягає 95 %).

Щоправда, необхідно зауважити, що дослідники хоча і використовували англомовні повідомлення, мали різні корпуси даних, що не дозволяє порівняти ці методи.

Кластеризація електронних повідомлень

Окрім класифікації повідомлень, певними авторами (G. Manco, E. Masciari, M. Ruffolo) для виявлення спаму були застосовані алгоритми видобування даних, зокрема кластеризація повідомлень [6]. Хоча ці автори використали різні типи інформації (структурну, числову тощо) та запропонували нову міру для визначення схожості повідомлень, загалом їхній підхід важко оцінити. Насамперед тому, що кожен користувач намагається впорядковувати свої повідомлення згідно зі своїми вподобаннями (себто, за тематикою, за відправником тощо), що ставить під сумнів практичну доцільність застосування цього підходу.

ПОСТАНОВКА ПРОБЛЕМИ

Розглянувши попередні дослідження, проблема була сформульована так:

Класифікувати німецькомовні повідомлення у дві категорії (спам – приватні повідомлення).

Вибір лише двох категорій зумовлений двома факторами:

З практичної точки зору користувач зацікавлений лише у виявленні та у видаленні спаму, тобто класифікація спаму у підкатегорії цікава лише з точки зору соціологічних досліджень.

Водночас, два різні користувачі здатні класифікувати одну множину повідомлень порізному, що робить недоцільним класифікацію приватних повідомлень у підкатегорії.

Класифікація з використанням підкатегорій доцільна у випадку використання поштової скриньки одного і того ж користувача. У цій статті розглядається змішаний корпус (дані від кількох користувачів)

Під час вибору мови повідомлень було враховано такі аспекти як

a) флективність мови;

b) кількість повідомлень, які було знайдено.

Після розгляду повідомлень шведською, польською та українською мовами, було обрано німецьку, оскільки вона задовольняла вимогу (a), а також було знайдено кілька джерел, які погодилися надати німецькомовний спам для цього дослідження.

РОЗПІЗНАВАННЯ СПАМУ ЯК ЗАДАЧА КЛАСИФІКАЦІЇ

Задачу навчання за зразками (базовану на попередньому досвіді) можна уявити так.

Маючи множину зразків (прикладів) для тренування (у даній статті об’єкт, зразок та приклад використовуються як синоніми (example), навчання за зразками або те, що базується на зразках/прикладах – example-based або memory-based learning)

$$T = \{(x_1, f(x_1)),...,(x_n, f(x_n))\},$$

де компоненти кожного прикладу x_i є векторами форми

$$x_i = (x_{i1}, x_{i2}, ..., x_{il}),$$

а $(x_{i1}, x_{i2}, ..., x_{il})$ – характеристиками (features або attributes) x_i, необхідно апроксимувати невідому функцію f, визначену на просторі входів Ω, у дискретний невпорядкований простір виходів $\{1, ..., K\}$. Об'єкти домена (зразки) представлені множиною пар характеристика-значення (feature-value) та класом, до якого вони належать.

Функція $f : \Omega \Rightarrow \{1, ..., K\}$ визначає К-частини простору входів як множини $f^{-1}(k)$, себто класи. Отримуючи на вхід множину зразків для тренування Т, алгоритм навчання продукує класифікатор , який є гіпотезою щодо функції f. Далі, отримуючи *x* нових значень, цей класифікатор передбачає відповідні значення *y*, тобто класифікує нові зразки.

У випадку розпізнавання спаму задача полягає у визначенні коректного класу/категорії (загальна кількість класів – 2 – спам та приватні повідомлення). Кожне повідомлення є окремим зразком/прикладом і наведено у формі характеристика-значення, де характеристиками у цьому конкретному випадку є текстові елементи повідомлення і значеннями – токен/слова.

ДАНІ

Повідомлення для дослідження були зібрані з кількох джерел, серед них з Центру опрацювання даних університету м. Тюбінген (K. Spanier), Штутґарту (U. Horlacher) та від кількох інших осіб.

Було відібрано 87 спам-повідомлень, а далі, розглянувши тематику цих повідомлень, 446 повідомлень з веб-форумів (фінансові повідомлення, а також на тему інформатики та відпочинку). Створений корпус містить повідомлення на різну тематику, щоб під час аналізу можна було оцінити, який тип повідомлень розпізнається найлегше та за яких умов.

Число повідомлень було обране з огляду на кількість та пропорцію повідомлень, використаних у інших дослідженнях. Так, кількість повідомлень у дослідженні Androutsopoulos-a є значно більшою, аніж у цьому, проте усі інші дослідження грунтувались на даних, де кількість спаму дорівнювала ста (або між 100 і 200) повідомленням [2].

ОПРАЦЮВАННЯ ДАНИХ

Дані опрацьовувалися так:

Токенізація.

Стемінг.

Створення списку найчастіше використовуваних слів.

Всі ці три етапи були вже описані вище. Однак необхідно наголосити, що при опрацюванні спам-повідомлень виникають певні труднощі, оскільки дуже часто слова в них розбиті пунктуаційними знаками (скажімо, ang*ebo*t – пропозиція) або ж розміщені вертикально (кожна буква слова в окремому рядку), що унеможливлює коректну токенізацію автоматичним шляхом.

На етапі стемінгу було використано модуль, написаний на Perl [4], котрий повертає слову "базову форму", себто інфінітив для дієслова, називний відмінок однини для іменників тощо. Якщо стеммер не може розпізнати слово, воно залишається у тій же ж формі. У багатьох випадках форма слова, яка є результатом виконання стемінгу, не є коректною, однак найголовнішим на цьому етапі є не правильний лінгвістичний аналіз, а

відображення різних форм одного й того ж слова в одній (тобто, якщо результатом стемінгу для слова Sie є si, що не є коректним з точки зору лінгвістики, але si є спільною формою для Sie у всіх відмінках, цей аналіз вважається достатнім). Водночас, якщо абсолютно різні слова отримують однакове відображення (wird->werden буде->бути, wir->werden ми->бути), це слід вважати за помилку. Тож цей простий лінгвістичний аналіз здійснюється з метою редукування кількості features і представлення різних форм слова як однієї.

Також під час опрацювання даних було створено список найчастіше вживаних слів для того, щоб видалити їх з корпусу повідомлень. Проте, побіжний аналіз спаму виявив, що займенники є важливою ознакою для виявлення німецькомовного спаму. Для прикладу, практично в усіх приватних повідомленнях використовується займенник die (ти), позаяк серед спаму найчастіше – Sie (Ви) (Du з'являється лише у спам-повідомленнях, що належать до категорії "флірт, знайомства"). Саме тому у список ввійшли лише артиклі і кілька прийменників (на відміну від англомовних списків).

ВИБІР ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ТА ПРЕДСТАВЛЕННЯ ДАНИХ

Для проведення цього дослідження було обране програмне забезпечення (ПЗ), створене Language Technology Group Університету Антверпен. Причиною використання цього ПЗ є розмаїття методів, представлених у ньому, зокрема, кількох методів присвоєння вагових коефіцієнтів характеристикам об'єктів, а також кількох метрик визначення схожості між характеристиками різних об'єктів.

Ще одним важливим чинником є те, як ці методи були зреалізовані. Оскільки підхід "лінивого навчання" (lazy learning), навчання за зразками передбачає збереження всіх прикладів у пам'яті, обчислювальна складність є пропорційною до числа зразків. Для роз'язання цієї проблеми розробники запропонували кілька оптимізацій, серед них представлення масиву даних у вигляді деревоподібної структури (tree-based memory) та інвертованого індексу (inverted index).

Враховуючи вибір TiMBL-програмного забезпечення для навчання та тренування, кожне повідомлення представлене у бінарному форматі як послідовність токенів, з класом, до якого це повідомлення належить як останній елемент цієї послідовності. Попередньо усі токени повідомлень були пронумеровані (тобто, якщо вони були представлені як 'ihre,gehen, ..., ihre, spam.', остаточний вигляд є '1,2,...,1,spam.'). Після стадії опрацювання даних було отримано 11821 токенів.

ТРЕНУВАННЯ ТА ТЕСТУВАННЯ

Найпростішою схемою визначення схожості між двома об'єктами є обчислення суми різниць між характеристиками цих об'єктів (алгоритм визначення k "найближчих сусідів" – k-NN – k-nearest-neighbours):

$$\Delta(X,Y)=\sum_{i=1}^{n}\delta(x_i,y_i)$$

де

$$\delta(x_i,y_i)=0 \;\; if \;\; x_i=y_i, \;\; \delta(x_i,y_i)=1 \;\; if \;\; x_i\neq y_i,$$

X,Y – об'єкти/зразки, $\Delta(X,Y)$ – відстань між ними, δ – відстань між і-ми характеристиками об'єктів.

Для тренування були сформовані такі схеми:

Схема	Алгоритм класифікації	Метод визначення вагових коефіцієнтів для характеристик	Тип “голосування” при виборі класу/категорії
1	IB2	Chi-squared feature weighting	Inverse linear weighting
2	IB2	Information-gain feature weighting (IG)	Inverse distance weighting
3	IB2	Gain Ratio weithting	Inverse distance weighting

де IB2 (incremental edited memory-based learning) – алгоритм, що дозволяє додавати до пам’яті ті зразки, які не були коректно класифіковані, оскільки вони потенційно є “найближчими сусідами” між собою.

Причиною обрання кількох методів визначення коефіцієнтів для характеристик є те, що хоча IG визначає, наскільки корисною є та чи інша характеристика з точки зору інформативності (IG характеристики *i* визначається як різниця ентропії у ситуації, коли значення *i* невідоме та тоді, коли відоме, Gain Ratio – нормалізоване IG), все ж було показано, що ці методи демонструють небажане зміщення в бік характеристик з більшою кількістю значень. Для запобігання цьому було запропоновано використовувати хі-квадрат критерій.

Під час вибору класу/категорії було вирішено керуватися двома методиками – інвертивної відстані та інвертивної лінійної схеми. Як альтернативою можна також користуватися голосуванням шляхом вибору більшості (majority voting), але небажано у випадку розріджених (sparse) даних.

У багатьох дослідженнях способом тренування/тестування було обрано валідацію типу 10-cross-validation, яка означає, що усі об’єкти поділено на групи по 10 у кожній, і кожна група є групою тестування, а інші – даними для тренування. Проте у випадку невеликої кількості даних краще здійснювати one-leave-out – кожен об’єкт підлягає тестуванню, а решта є тестувальним матеріалом.

РЕЗУЛЬТАТИ ТА ОЦІНКА

Схема	Категорія	Точність	Повнота
1	Спам	100,00 %	64,37 %
	Приватні повідомлення	94,18 %	100 %
	Разом (точність)	93,50 %	100 %
2	Спам	100,00 %	74,71 %
	Приватні повідомлення	95,30 %	100,00 %
	Разом (точність)	95,87 %	
3	Спам	100,00 %	74,71 %
	Приватні повідомлення	95,30 %	100,00 %
	Разом (точність)	95,87 %	

Критеріями оцінки ефективності є точність та повнота (precision – recall). Точність обчислюється як частка правильно класифікованих повідомлень серед усіх повідомлень, що класифіковані до цієї категорії. Повнота є часткою правильно класифікованих повідомлень серед усіх повідомлень, що справді належать до цієї категорії.

Точність для обидвох категорій у найкращому випадку становить 95,87 %, що відповідає результатам Androutsopoulos-a [1] (96,89 %). Проте, хоча він теж використовував навчання, що базується на прикладах, кількість характеристик у його дослідженні становила від 50 до 300 (у цьому – більш, аніж 11000). На жаль, неможливі прямі порівняння з іншими дослідженнями, оскільки (а) для них використовувалися різні множини даних; (б) різні методи; (с) різні мови (англійська – німецька).

Важливим результатом проведення цього експерименту є те, що жодне приватне повідомлення не було класифіковане як спам (повнота для категорії "приватні повідомлення" становить 100 %). Androutsopoulos [1] у своєму дослідженні запропонував використовувати різні вагові коефіцієнти під час оцінки результатів, оскільки кошт класифікації приватних повідомлень як спаму є значно вищим, аніж присвоєння спам-повідомленню категорії "приватні повідомлення".

Як показано на рисунку, результати з використанням схеми 2 та 3 є однаковими. Однак, використання критерію хі-квадрат лише знизило повноту для спам-повідомлень. Це, вочевидь, можна пояснити тим, що цей розподіл слід використовувати у випадках, коли очікувана частотність більш аніж 20 % токенів є меншою, ніж 5 або хоча б 1 % менша аніж 1. Як описано вище, використання стемера на етапі опрацювання даних не завжди редукувало форми одного й того ж слова, що відповідно проявилося у виникненні надлишкових токенів; окрім того, бажання використати спам різного типу привело до розрідженості даних.

Також слід наголосити: однією зі специфік використання підходу, який ґрунтується на навчанні з використанням попередніх зразків, є те, що якщо новий об'єкт/приклад не з'являвся раніш у множині даних для тестування, є мала ймовірність, що він буде класифікований коректно.

ВИСНОВКИ

Результати цього дослідження підтвердили тезу, що застосування підходу "машинного навчання", зокрема, одного з типів – memory-based learning, конкурує з використанням підходу Баєса. Точність та повнота, отримані як для спаму, так і для приватних повідомлень німецькою мовою, є достатньо високими і відповідають результатам, що досягнуті для англійської мови. Автор сподівається, що результати можуть бути покращені, якщо використати додаткову інформацію як характеристики повідомлень. Окрім того, дослідження, що виконуються з використанням поштової скриньки лише одного користувача, повинні підвищити рівень повноти для спаму (враховуючи те, що користувач часто отримує спам однієї і тієї ж тематики або кілька ідентичних спам-повідомлень, є добрим аргументом для використання навчання, яке грунтується на прикладах). Ще один шлях до покращення результату – це редукування надлишкових токенів та використання кращого засобу для здійснення стемінгу.

1. Androutsopoulos, I., G. Paliouras, V. Karkaletsis, та інші. Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and A Memory-Based Approach. Proceedings of the Workshop "Machine Learning and Textual Information Access", European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). – Lyon, France, 2000. –

P. 1–13. 2. Crawford E., J. Kay, E. McCreath. Automatic Induction of Rules for e-mail Classification. Proceedings of the Sixth Australasian Document Computing Symposium, Coffs Harbour, Australia, December 7, 2001. 3. Daelemans W., J. Zavrel, K. van der Sloot, and Antal van den Bosch. TiMBL: Tilburg Memory Based Learner, version 4.2, Reference Guide. ILK Technical Report 02-01, Available from http://ilk.kub.nl/downloads/pub/papers/ilk0201.ps.gz, 2002. 4. German stemmer. http://search.cpan.org/author/ULPFR/Text-German-0.03/German.pod . 5. Hedstroem A., S. Katrenko, Ch. Larsson. Filtering Spam Mail Using Memory-Based Learning (TiMBL). Presentations in Information Retrieval. Students Workshop in Tübingen, May 29th – 31th, 2003. 6. Manco G., E. Masciari, M. Ruffolo, A. Tagarelli. Towards An Adaptive Mail Classifier. 7. Pantel P., D. Lin. SpamCop: A Spam Classification & Organization Program. In Proceedings of AAAI-98 Workshop on Learning for Text Categorization. – P. 95–98. Madison, Wisconsin, 1998. 8. Provost J. Naive-Bayes vs. Rule-Learning in Classification of Email. University of Texas at Austin, Artificial Intelligence Lab. Technical Report AI-TR-99-284.

УДК 518

Я.П. Кісь, О.Я. Тарас
Національний університет “Львівська політехніка”,
кафедра “Інформаційні системи та мережі”

ОСОБЛИВОСТІ ПРЕДСТАВЛЕННЯ ГРАФІЧНОЇ ІНФОРМАЦІЇ В INTERNET

The format of the graphics files and the specifics of there use for the web-systems are represented in this article.

Розглянуто формати графічних файлів та особливості їх застосування для представлення графічної інформації у web-системах

ВСТУП

Поговоримо про графіку на Web-сторінках, адже саме завдяки їй WWW став найпопулярнішим сервісом Internet, саме їй ми зобов’язані розмаїттю інформації. Постає природне запитання: що є особливого в графіці, яка застосовується на web-сторінках? Відповідь проста – вона має свої певні обмеження, які ми повинні враховувати з максимальною вигодою для себе. Для розробки web-сторінок використовуються два основні формати файлів – GIF і JPG. Зараз з’явився новий формат для web-графіки за назвою PNG (вимовляється “пинг”), але він поки ще мало поширений, і не всі броузери його розуміють, тому про нього згадувати не будемо. Опишемо основні властивості й особливості форматів GIF і JPG.

АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ

Всесвітня інформаційна мережа надає унікальні можливості передачі інформації і проведення направлених маркетингових досліджень. Ці задачі неможливо вирішити без використання графіки. Хоча графіка не замінить собою текст, але, безумовно, вона зможе вигідно доповнити і проілюструвати його. Адже графіка іноді містить інформацію, яку