

Тарасенко В.П., д.т.н., проф., зав. каф. СКС НТУУ „КПІ” (м.Київ)

Михайлюк А.Ю., к.т.н., доц. каф. СКС НТУУ „КПІ” (м.Київ)

Тесленко О.К., к.т.н., доц. каф. СКС НТУУ „КПІ” (м.Київ)

Осипов О.С., аспірант каф. СКС НТУУ „КПІ” (м.Київ)

АВТОМАТИЗАЦІЯ ОЦІНКИ ОРИГІНАЛЬНОСТІ ІНФОРМАЦІЇ

Тарасенко В.П., Михайлюк А.Ю., Тесленко О.К., Осипов О.С.

Автоматизація оцінки оригінальності інформації.

В напрямку підвищення стійкості інформаційних технологій розглядаються можливі підходи до формалізації поняття оригінальності інформації з метою автоматизації процесу визначення кількісних оцінок. Пропонується нова інформаційна технологія ієрархічного адаптивного порівняння, яка має ряд суттєвих переваг перед відомими засобами визначення запозичень. Запропоновані підходи до формалізації поняття оригінальності інформації дозволяють автоматизувати процес визначення відповідних кількісних оцінок, що з розвитком комп'ютерних мереж та інформаційних технологій має велике практичне значення.

Tarasenko V.P., Mykhailyuk A.Y., Teslenko O.K., Osypov O.S.

An automation of information originality assessment.

In the field of improvement of information technologies withstandability the approaches to information originality formalization are evaluated in the context of automation of information originality quantitative assessment. The new information technology of hierarchical adaptive comparison is proposed, with a set of distinct advantages over an existing borrowing detection means. Developed approaches to information originality formalization allow automation of its quantitative assessment that in context of rapid evolution of information technologies and networks has a great practical value.

Вступ

З розвитком та широким впровадженням комп'ютерних технологій та мереж все більшу актуальність набувають дослідження та розробки в галузі їх інформаційної стійкості [1]. Проблематика інформаційної стійкості, маючи споріднену природу із проблематикою гарантоздатності [2], потребує врахування не лише аспектів надійності та безпеки інформації, але й більш повного врахування когнітивних властивостей інформації для забезпечення протидії негативним антропогенним впливам, зокрема нештатному використанню комп'ютерних систем. Прикладами когнітивних властивостей інформації є цілісність, конфіденційність, унікальність та оригінальність [1]. При цьому проблематика цілісності та

конфіденційності в таких, наприклад, напрямках як завадостійке кодування та криптографічні перетворення, досліджена значно глибше ніж проблематика унікальності та оригінальності. Це пояснюється пріоритетами суспільної практики на відповідний момент часу. Властивості унікальності та оригінальності інформації формують основу оцінки когнітивної цінності інформаційних об'єктів та певний базис оцінки належності інформаційного об'єкта відповідному користувачу комп'ютерної системи чи іншому суб'єкту віртуального простору. Наприклад, в контексті освітніх комп'ютерних технологій забезпечення властивості унікальності інформації є бажаним, а оригінальності обов'язковим. Тому в сучасних умовах суспільна практика вже гостро потребує, особливо в науково-освітній галузі, конкретних результатів досліджень проблематики оригінальності інформації [3,4,5]. Одна з першочергових проблем полягає в формалізації поняття оригінальності до рівня, достатнього для його використання в інформаційних технологіях.

Формалізація поняття оригінальності на прикладі освітньої галузі

В реальній освітній практиці значна увага приділяється самостійному вирішенню студентом (учнем, абітурієнтом) одержаного завдання. При прийнятті робіт фактично виконуються дві перевірки – перевірка правильності виконання завдання та перевірка оригінальності, тобто завдання виконано самостійно чи вирішення завдання запозичено. Якщо в першому випадку перевірка здебільшого може бути виконана без присутності автора, то в другому випадку в традиційних освітніх технологіях для перевірки рівня оригінальності необхідний безпосередній контакт з автором для проведення діалогу в режимі реального часу. В окремих випадках, наприклад на вступних іспитах, для забезпечення оригінальності використовуються жорсткі організаційні заходи.

З розвитком інформаційних технологій з'явилась можливість виконувати запозичення практично без особливих зусиль. Якщо у доінформаційну епоху необхідно було, принаймні, самому переписати вирішення завдання (і тим самим хоча б на рівні підсвідомості ознайомитись з його змістом), то нині переписування успішно і без помилок виконує комп'ютер. Тобто можливості запозичень незрівнянно зросли, а необхідність особистої участі в переписуванні чужого матеріалу при запозиченні зменшилась. З іншої сторони розвиток комп'ютерних мереж та технології дистанційного навчання зменшили необхідність безпосереднього контакту між викладачем і студентом в режимі реального часу. Звідси впливає значне посилення загрози невиявленого запозичення результатів виконання робіт, а відповідно, й практична необхідність автоматизації оцінювання оригінальності інформації.

Згідно з [5] оригінальність – властивість інформації нести відбиток творчих та особистісних якостей автора, що в освітній сфері є запорукою самостійного виконання завдання студентом.

Одним із варіантів такого відбитку є стиль. У вузькому розумінні стиль – це сукупність особливостей у побудові мови (тексту), манера словесного викладу. У широкому – сукупність характерних мовних та граматичних особливостей, особливостей аналітичних та логічних суджень, рівня розуміння та сприйняття термінології предметної галузі, культурного рівня та навіть власних уподобань, що безпосередньо впливають на інформаційні об'єкти, створені людиною. Мірою оригінальності інформації в цьому разі виступає величина, яка характеризує рівень співпадіння стилю піддослідного документу зі стилем, притаманного особі, яка видає себе за автора.

В суто текстових документах (рефератах, пояснювальних записках, доповідях та інших), формальними характеристиками стилю може виступати використання типових слів та словосполучень, певної термінології та сленгу, сукупний словниковий запас особистості, типові, для особистості, синтаксичні та граматичні помилки. Маючи певну, відносно малу, базу текстових документів, створених студентом (учнем, абітурієнтом), можна виділити сукупність формальних характеристик його стилю, за згаданими вище ознаками, та використовувати їх для визначення оригінальності його майбутніх робіт. Зрозуміло, що процес навчання, розвитку та самоствердження особистості знаходить своє відображення у поступовій зміні її стилю, а отже й сукупність формальних характеристик її стилю не можна вважати статичною. Сукупність характеристик стилю певної особи необхідно доповнювати й переоцінювати не лише з урахуванням нових документів (інформаційних об'єктів), а й поступово зменшуючи вплив старих документів на ці характеристики.

Для більш „когнітивно складних” інформаційних об'єктів, наприклад, різноманітних обчислювальних та дослідницьких робіт, вихідних кодів програм, креслень та малюнків можна виділити й інші додаткові формальні характеристики стилю. Можливість визначення повної сукупності таких характеристик значною мірою залежить від специфіки навчального закладу, специфіки галузі, способів та засобів перевірки знань студентів, наявності відповідної технічної бази та кваліфікованих фахівців для проведення аналізу.

Однак, незважаючи на проблеми, пов'язані з повнотою формального відображення стилю, такий підхід в визначенні оригінальності інформації має суттєву перевагу - порівняно незначний обсяг даних, які необхідно обробляти, оскільки ці дані зосереджені в піддослідному документі та в попередньо сформованій характеристиці стилю особи, яка видає себе за автора. Але враховуючи існуючий рівень знань про процеси формування даних в свідомості людей, співпадіння стилю може з достатньою

достовірністю визначати оригінальність підслідного документу, але відсутність такого співпадіння не може бути достовірною ознакою використання автором запозичень. Відповідно, прийняття адміністративних заходів впливу в цьому випадку може бути оскаржене. Таке оскарження неможливе, якщо будуть надані документи, з яких взяті запозичення. Звідси випливає необхідність дещо іншого підходу до формалізації визначення оригінальності інформації в електронних документах, а саме - як міри відсутності запозичень з інших електронних документів. Такий підхід дозволяє документально довести наявність запозичень та прийняття відповідних адміністративних заходів впливу. Але достовірність визначення оригінальності підслідного документу залежить від кількості документів, з якими проводилось порівняння, а така кількість в загальному випадку практично нескінчена.

Таким чином, розглянуті підходи до формалізації властивості оригінальності вдало доповнюють один одного - в першому випадку з достатньою достовірністю можна визначити оригінальність інформації, а в другому випадку достовірно визначити її неоригінальність, що відповідає комбінованому методу визначення цієї когнітивної властивості інформації [6].

Використання існуючих інформаційних технологій для здійснення несанкціонованих запозичень та їх приховування, вимагає створення контролюючих інформаційних технологій для відповідної протидії. Базуючись на формалізації оцінок властивості оригінальності, такі технології повинні задовольняти наступним вимогам:

1. Подолання існуючих методів приховування запозичень від візуального контролю перевіряючих (переформатування, зміна заголовків, об'єднання абзаців та розбивка існуючих, видалення або перестановка перших та останніх абзаців в розділах та підрозділах і т.п.).

2. Подолання можливих методів приховування запозичень від контролюючих інформаційних технологій (наприклад, глобальне введення незначущих додаткових символів, глобальна заміна на букви латинського алфавіту, які мають те ж зображення, глобальна заміна на слова-синоніми і т.п.).

3. Витрати при запозиченнях (наприклад, особистого часу) на подолання можливостей контролюючих інформаційних технологій повинні бути значно більшими за витрати на самостійне вирішення завдання.

Очевидно, що наведені вимоги стосуються перед усім оцінки оригінальності на основі порівняння. Оцінка оригінальності на основі стилю дозволяє відмовитись від порівнянь, якщо інформація в електронному документі визначена як оригінальна і тим самим зменшити загальний обсяг порівнянь. Для подальшого зменшення обсягу порівнянь необхідно мінімізувати кількість документів з яких, можливо, взяті

запозичення. В деякій мірі можуть бути використані адміністративні заходи в межах відповідного інформаційного середовища. Наприклад, в межах ВНЗ це може бути організація електронних архівів студентських робіт на кафедрах. Але навіть в цьому випадку мінімізація кількості документів для порівнянь буде неоптимальною, оскільки з одного боку навіть в межах кафедри студентські роботи можуть значно відрізнятися по своїй тематиці, а з іншого боку різні кафедри можуть мати однакову тематику робіт. Крім того, наявність та доступність глобальної комп'ютерної мережі (Internet) вносить додаткові складнощі.

Суттєве звуження кола підконтрольних документів може бути досягнуто шляхом залучення інформаційно-пошукових систем. Однак, їх використання для підбору потенційних джерел запозичень пов'язано з деякими складнощами. Зокрема це викликано тим, що переважна більшість наявних на сьогоднішній день реально доступних „пошуковиків” орієнтована перш за все на контекстний (ключовий) пошук. Оскільки ж потенційно запозичені сегменти підконтрольного тексту локалізувати принципово не завжди можливо, ефективність застосування пошукового інструментарію виявляється досить низькою. Істотні переваги в цій ситуації міг би надати пошук за змістом, однак можливість семантичного пошуку інформації функціональністю сучасних „загальнодоступних” Internet-пошуковиків, як правило, не передбачена. Тому для знаходження потенційних джерел запозичень може бути застосований опосередковано семантичний пошук з використанням у якості критерію ключових термінів, котрі характеризують відповідну тематику. Однак, як показує досвід, пошук за розширеним набором ключових термінів призводить до розростання масиву результату пошуку за рахунок документів, котрі є релевантними, тобто формально відповідають набору пошукових ключів, але не є пертинентними, оскільки не задовольняють реальних потреб користувача. У якості ефективних засобів відфільтровування пошукового шуму можуть бути використані засоби кластеризації документів за ознакою тематичної належності.

Вказані засоби кластеризації та відповідні адміністративні заходи в межах єдиного інформаційного середовища ВНЗ дозволять значно скоротити практично необмежену кількість електронних документів, які використовуються для порівнянь до прийняттого, хоча і значного, рівня.

Технологія ієрархічного адаптивного порівняння

Інформація, яка циркулює в сучасних комп'ютерних системах та мережах, в загальному випадку має досить складну ієрархічну структуру з наявністю типових фрагментів. В зв'язку з цим на першому (нижньому) рівні ієрархії інформацію в електронному документі можна подати як неструктуровану послідовність символів деякого алфавіту, наприклад, як послідовність байтів. Інші рівні ієрархії визначають когнітивний зміст

інформаційних об'єктів – текстовий документ, графічні об'єкти, аудіо та відео фрагмента та інші.

Будемо розрізняти послідовність символів S , яка перевіряється (тобто, для якої обчислюється значення оригінальності, піддослідна послідовність) та послідовність символів E_i , ($i=1,2,\dots,n$), n – кількість електронних документів, визначених для пошуку можливих запозичень (послідовностей-зразків). Запозиченням будемо вважати будь-яку підпослідовність (довільний суцільний фрагмент послідовності) символів алфавіту довжиною **не менше** d символів, де d достатньо більше за 1, яка міститься як в піддослідній послідовності, так і в послідовності-зразку. Значення d будемо називати мінімальною довжиною запозичення, яка береться до уваги. Всі символи послідовності S , які містяться в тому чи іншому запозиченні, відмічаються. Відносна кількість (наприклад, в процентному відношенні) невідмічених символів i є мірою оригінальності послідовності S . Необхідно зауважити, що символи послідовності S , відмічені при порівнянні з послідовністю E_a , не можуть бути видалені при порівнянні з послідовністю E_b ($a \neq b$, $a, b = 1, 2, \dots, n$), оскільки можуть бути втрачені запозичення довжиною d і більше символів із послідовності E_b (при наявності певних співпадінь в послідовностях E_a та E_b). В такому разі може бути використано одночасне порівняння послідовності S з багатьма послідовностями-зразками з наступним накладенням результатів порівнянь для обчислення рівня оригінальності, тобто можливе розпаралелювання процесу обчислень.

Позначимо далі через c ($1 \leq c < d$) мінімальну довжину (в символах) початкового порівняння. Виберемо із послідовності S з кроком x , а із послідовності $E \in \{E_1, E_2, \dots, E_n\}$ з кроком y підпослідовності довжиною c . Величину x виберемо так, щоб будь-яка підпослідовність із S з мінімальною довжиною запозичення (d) повністю містила одну із вибраних підпослідовностей довжиною c , тобто $1 \leq x < d - c$. Позначимо S_j ($j=1,2,\dots,r$) вказані підпослідовності із S , $r=(L_S-(d-c)) \text{ div } x$, L_S – довжина послідовності S . Позначимо, також, F_h ($h=1,2,\dots,t$) вказані підпослідовності із E , $t=(L_E-c) \text{ div } y$, L_E – довжина послідовності E .

Розглянемо наступний базовий алгоритм технології. Визначення рівня оригінальності S по відношенню до E виконується в два етапи – етап швидкісного і етап детального порівняння. На етапі швидкісного порівняння кожна підпослідовність S_j порівнюється з кожною підпослідовністю F_h , і в випадку хоча б одного співпадіння відповідна підпослідовність S_j відмічається. Якщо в результаті для деякого a ($a \in \{1, 2, \dots, r-1\}$) S_a та S_{a+1} не відмічені, то це означає, що всі символи між цими підпослідовностями, а також символи самих підпослідовностей (всього $x+c$ символів), при умові, що $y=1$, гарантовано не можуть бути запозиченням із E при заданій мінімальній довжині запозичення. Таким чином етап швидкісного порівняння визначає рівень оригінальності S по

відношенню до E , який не може зменшитись при більш детальних дослідженнях. На етапі детального порівняння в кожному випадку співпадінь підпоследовностей S_j та F_h виконується посимвольне порівняння з місця виявленого співпадіння, як в напрямку закінчення так і в напрямку початку последовностей S та E , на кількість символів, яка не перевищує d . У випадку, коли буде виявлена спільна підпоследовність довжиною d і більше символів, то відповідні символи последовності S відмічаються. Легко перевірити, що в результаті детального порівняння в последовності S будуть відмічені всі символи, які містилися принаймні в одній спільній з E підпоследовності довжиною не менше d .

Базовий алгоритм технології дозволяє на швидкісному етапі визначити нижню границю оригінальності (або верхню границю відносного рівня запозичень), а на етапі детального порівняння – точне значення оригінальності (точний рівень запозичень, точний рівень співпадінь).

Складність $C_{ш}$ швидкісного порівняння по кількості операцій посимвольних порівнянь можна точно визначити як $r \times t \times c$, або

$$C_{ш} \approx L_S \times L_E \times c / x \quad (1)$$

Максимальна складність C_{∂} етапу детального порівняння визначається наступним чином:

$$C_{\partial} \leq K_c \times 2 \times (d - c),$$

де K_c кількість співпадінь підпоследовностей S_j та F_h .

Із одержаних значень $C_{ш}$ та C_{∂} впливає наступне. Величина x не впливає на C_{∂} , тому її значення необхідно вибирати максимально можливим, а саме: $x = d - c - 1$. Подальше збільшення x пов'язане зі збільшенням мінімальної довжини запозичення, яка береться до уваги. В свою чергу величину d доцільно установлювати (адаптувати) автоматично або вручну в залежності від конкретних умов визначення запозичень. Наприклад, значення d можна автоматично встановлювати пропорційно значенню L_S . В цьому випадку $C_{ш}$ не буде суттєво залежати від довжини последовності, яка перевіряється. Збільшення величини c призводить до збільшення $C_{ш}$ і до зменшення C_{∂} , тобто виникає протиріччя. Зменшення C_{∂} виникає, перед усім, за рахунок зменшення K_c , оскільки при збільшенні мінімальної довжини початкового порівняння зменшується ймовірність випадкових співпадінь. Для усунення протиріччя можна використовувати як апаратні так і програмні засоби. В першому випадку на апаратні засоби покладається виконання одноктного багатосимвольного порівняння. Наприклад, якщо в якості символу вибрано один байт, то в сучасних мікропроцесорах існують команди одночасного порівняння до 8 байт. Тобто в формулі (1) значення c може бути прийнято за 1, але значення K_c буде формуватись при значно меншій ймовірності випадкових співпадінь 8-байтних підпоследовностей, порівнюючи з ймовірністю співпадінь окремих байтів. В другому випадку початкові порівняння

підпоследовностей із $c \gg 1$ символів трансформуються в порівняння відповідних дайджестів цих підпоследовностей. При цьому виникає проблема формування оптимальних алгоритмів їх обчислення.

Подальший розвиток базового алгоритму технології полягає у використанні спеціалізованих процедур (пристроїв) для узагальнення початкових порівнянь на випадок $y > 1$ і встановлення в результаті значень C_{in} та C_o , в яких коефіцієнт пропорційності відносно L_E був би незначним.

В технології ієрархічного адаптивного порівняння базовий алгоритм нижнього рівня ієрархії використовується для порівнянь "символів" в базовому алгоритму більш високого рівня ієрархії – тобто використовується своєрідна рекурсія. Враховуючи, що практично всі електронні документи мають складну ієрархічну структуру, елементи нижнього рівня ієрархії можна розглядати як „символи” для базового алгоритму верхнього рівня ієрархії. Порівняння таких “символів” виконується за допомогою базового алгоритму нижнього рівня ієрархії з урахуванням можливої автоматичної або автоматизованої адаптації значень як мінімальної довжини запозичення так і рівня оригінальності, при якому “символи” вважаються неспівпадаючими. Наприклад, у випадку текстових документів типовими можна вважати наступні рівні ієрархії – символи – речення – абзаци – підрозділи - розділи. На самому нижньому рівні ієрархії за допомогою базового алгоритму порівнюються окремі речення. При цьому, завдяки адаптивному визначенню мінімальної довжини запозичення та допустимому проценту запозичень, співпадаючими можуть бути визначені речення, які відрізняються між собою шляхом видалення, вставки, заміни окремих слів та словосполучень. На більш високому рівні ієрархії можуть бути визнані співпадаючими, наприклад, підрозділи тексту, які відрізняються між собою шляхом видалення, вставки, заміни окремих речень чи абзаців. Тобто технологія ієрархічного адаптивного порівняння може визначити наявність запозичень при досить глибокому їх маскуванні. Звичайно, остаточний вердикт - чи це замасковане запозичення, чи усвідомлене самостійне викладення матеріалу, в випадку прийняття адміністративних заходів, може виконати лише людина.

На кафедрі спеціалізованих комп'ютерних систем НТУУ "КПІ" проведена розробка прототипу комплексу програм, який реалізує основні аспекти технології ієрархічного адаптивного порівняння. Дослідна експлуатація комплексу підтвердила високу ефективність запропонованої технології.

Висновки

Запропоновані підходи до формалізації поняття оригінальності інформації дозволяють автоматизувати процес визначення відповідних

кількісних оцінок, що з розвитком комп'ютерних мереж та інформаційних технологій має велике практичне значення.

Запропонована технологія ієрархічного адаптивного порівняння для визначення рівня оригінальності на основі порівняння має наступні позитивні якості: простота програмної реалізації завдяки рекурсивним властивостям базового алгоритму; можливість автоматичної або автоматизованої адаптації до конкретних умов; можливість досягнення досить значної швидкодії, що зокрема забезпечується можливістю незалежності швидкодії від довжини послідовності, яка перевіряється та можливістю розпаралелювання процесу обчислень на будь-якому із рівнів ієрархії; можливість застосування технології для порівняння текстових документів на будь-яких однакових мовах, вихідних кодів програм та інших інформаційних об'єктів.

Оскільки в науково-освітній сфері унікальність інформації асоціюється з таким поняттям, як наукова новизна, то унікальна інформація в цій сфері не може не бути неоригінальною, тому результати визначення оригінальності можуть розглядатись як початковий етап автоматизації визначення унікальності інформації.

Незважаючи на те, що ефективність отриманих результатів розглядалась на прикладі освітньої сфери, вони можуть бути поширені в цілому ряді інших галузей, де важливим є визначення подібності складних ієрархічних структур (контроль інтелектуальної власності, контроль технологічних процесів, контрольно-ревізійна служба, аналіз ДНК та інші).

Подальший розвиток теоретичних основ та практичних засобів автоматизації оцінки унікальності та оригінальності доцільно проводити в наступних напрямках: створення інформаційної технології і відповідного програмного забезпечення для автоматизації визначення кількісних характеристик стилю; розробка та удосконалення семантично орієнтованих інформаційно-пошукових систем та програмних засобів тематичної кластеризації результатів пошуку; розвиток формалізації поняття оригінальності для випадків, коли як відсутнє співпадіння стилю так і відсутні виявлені запозичення; проведення статистичних досліджень технології ієрархічного адаптивного порівняння для повноти автоматичного визначення параметрів адаптації; створення спеціалізованих операційних пристроїв та спеціалізованих співпроцесорів з використанням технології ПЛІС для підвищення ефективності базового алгоритму технології.

Література

1. Тарасенко В.П., Михайлюк А.Ю., Тесленко О.К., Осипов О.С. Методологічні та термінологічні аспекти інформаційної стійкості освітніх комп'ютерних технологій та мереж // Радіоелектронні та комп'ютерні системи.- 2006. - №7, - с. 28-31.

2. Харченко В.С. Гарантоздатність комп'ютерних систем: проблеми і результати // Авіаційно-космічна техніка і технологія.- 2005.- №7(23).- С.352-376.
3. <http://www.antiplagiat.ru>
4. <http://www.TurnItIn.com>
5. Tarasenko V.P., Mykhailyuk A.Y., Teslenko O.K., Osypov O.S. The Information Withstandability of the Educational Computer Technologies and Networks // In Proceedings of the Advanced Computer Systems and Networks: Design and Application – Ukraine, Lviv, September 21-23, 2005.
6. Тесленко А.К., Осипов А.С. Информационная стойкость: уникальность и оригинальность информации. Тезисы 7-ой международной научно-практической конференции «Современные информационные и электронные технологии», Одесса, 2006.