УДК 004.021

О. В. Бісікало, Г. О. Кириленко
Вінницький національний технічний університет

МЕТОД ФОРМАЛЬНОГО ВИЗНАЧЕННЯ ЯКОСТІ ДОПИСУ НА СПЕЦІАЛІЗОВАНИХ САЙТАХ

Розглянуто метод формального визначення якості посту на основі множини вибраних параметрів. Для вирішення цієї проблеми пропонується застосувати Java-бібліотеку Jsoup для парсингу HTML-коду, а також засоби Matlab для побудови дерева рішень, що використовується для визначення показника якості посту.

Ключові слова: пост, парсинг, Jsoup, Matlab, дерево рішень.

Post quality assessing algorithm based on the set of chosen parameters is considered in the article. To solve the problem the following next instruments will be used: Java library called Jsoup for HTML-code parsing, and Matlab tools for building the decision tree for post quality assessing.

Key words: post, parsing, Jsoup, Matlab, decision tree.

Вступ

Сьогодні ми все частіше використовуємо Інтернет для пошуку потрібної інформації. Зокрема, багато корисної інформації є на форумах чи інших спеціалізованих сайтах. Проте на пошук корисних дописів, які все частіше називають запозиченим з англійської словом *пост*, зазвичай доводиться витрачати надто багато часу. Тому актуальною є проблема автоматичної оцінки якості допису (посту), що значно зменшить час пошуку потрібної інформації [1–3].

Аналіз останніх досліджень і публікацій

Проблему автоматичного визначення якості допису розглянуто в праці «Automatically Assessing the Post Quality in Online Discussions on Software» [4] вчених лабораторії обробки знань відкритого доступу університету Дармштадта. Вони запропонували декілька категорій параметрів дописів – лексичні, граматичні тощо. Для отримання результатів тут використовується машинне навчання, найкращі результати – до 90 % правильних оцінок допису.

Формування мети

Мета дослідження – обґрунтувати інформативні параметри допису, які впливають на його якісний показник корисності, а також розробити формальний метод автоматичного визначення якості допису на основі вибраних параметрів.

Аналіз отриманих наукових результатів

Сьогодні дуже популярний серед розробників програмного забезпечення сайт http://stackoverflow.com/ [5]. Тут користувачі можуть залишати свої питання, а також відповідати на питання інших чи коментувати їх. Одне запитання може мати багато відповідей. Користувачі оцінюють їх корисність своїми голосами. Чим більше голосів набрала відповідь, тим вона краща. Корисних відповідей може бути декілька, причому корисними в цьому випадку вважаються ті відповіді, які набрали більше голосів порівняно з рештою. Спочатку спробуємо визначити, які параметри впливають на якість відповіді [6, 7]. Автоматизація змушує, насамперед, визначити –

як зрозуміти, чи відповідь корисна, не аналізуючи її змістове наповнення? На початку дослідження припускаємо, що для визначення якості допису мають значення такі формальні параметри:

1. Перше – це розмір допису-відповіді. Занадто короткі відповіді можуть не повністю розкривати питання, а занадто довгі апріорі можуть містити зайву інформацію.

2. Друге – це наявність посилань. Зазвичай посилання на інші джерела – часто це книги, документації – бувають дуже корисними.

3. Третім параметром є наявність знаків питання. Якщо автор відповіді задає питання, тобто не до кінця зрозумів питання або уточнює певні деталі, це дає змогу припустити, що відповідь не повна.

4. Четвертий параметр – наявність трьох крапок, що часто свідчить про сумніви автора щодо правильності своєї відповіді.

5. Останнім параметром є відсоток коду у відповіді. Оскільки цей сайт є спільнотою розробників програмного забезпечення, то часто питання і відповіді містять програмний код, без якого відповідь часто неможлива.

Усі перераховані параметри обґрунтовано на основі власного досвіду користування цим сайтом. У цьому дослідженні перевіримо, чи є інформативними вибрані параметри.

Для отримання цих ознак будемо парсити html-код сторінки за допомогою java-бібліотеки jsoup, яка містить багато готових функцій для отримання певних тегів зі сторінки, тексту між тегами тощо. Спочатку ми отримуємо масив відповідей, а потім – їх окремі параметри. Одержані значення параметрів записуємо у файл. Нижче показано фрагмент вихідного файла, де кожен рядок – це допис-відповідь. У стовпцях записано параметри допису. Останнім параметром є результат – корисний допис чи ні. «1» означає, що допис корисний, «2» – не корисний. Цей параметр визначено на основі відданих за відповідь голосів. Якщо багато користувачів проголосували позитивно за відповідь, то вона вважається корисною.

Таблиця 1

Вихідний файл

Розмір допису	Кількість коду у дописі	Кількість «?»	Кількість посилань	Наявність «…»	Чи корисний допис?
X1	**X2**	**X3**	**X4**	**X5**	**1 – так, 2 – ні**
1	3	0	0	0	1
1	1	0	0	0	1
4	5	0	0	0	1
1	2	0	1	0	2
1	4	0	0	0	2
2	4	0	0	0	2
5	1	1	0	0	2
…	…	…	…	…	…

Для визначення параметрів використаємо дерева рішень, для побудови яких застосуємо пакет прикладних програм MatLab [8–10]. Дерева рішень будуються автоматично, тому вихідне дерево використовує лише інформативні параметри. Побудувавши дерево рішень, побачимо, які з запропонованих параметрів мають значення для автоматичного визначення якості допису.

Вхідні змінні можуть бути задані у вигляді неперервних величин або категорій. У ході дослідження випробувано декілька способів подання таких параметрів, як розмір допису і відсоток коду. Спочатку їх задавали кількістю символів. Проте з такими вхідними даними згенероване дерево рішень було завелике. Тому запропонували задати вибрані параметри в категоріальному вигляді. Нижче у табл. 2 наведено пояснення всіх вхідних параметрів для побудови дерева рішень.

Спочатку будуємо дерево на основі параметрів всіх дописів. Таке дерево дає рішення на основі всіх запропонованих параметрів, з чого можемо зробити висновок, що всі запропоновані параметри інформативні.

Таблиця 2

Вхідні параметри для побудови дерева рішень

Змінна	Значення
X1 – Розмір допису	1 – розмір допису становить від 0 до 20 відсотків від розміру найбільшого допису на сторінці 2 – 21–40 % 3 – 41–60 % 4 – 61–80 % 5 – 81–100 %
X2 – Кількість коду у дописі	1 – кількість коду становить від 0 до 20 відсотків від загального розміру допису 2 – 21–40 % 3 – 41–60 % 4 – 61–80 % 5 – 81–100 %
X3 – Кількість «?»	1..∞
X4 – Кількість посилань	1..∞
X5 – Наявність «…»	1 – містить «…» 2 – не містить «…»

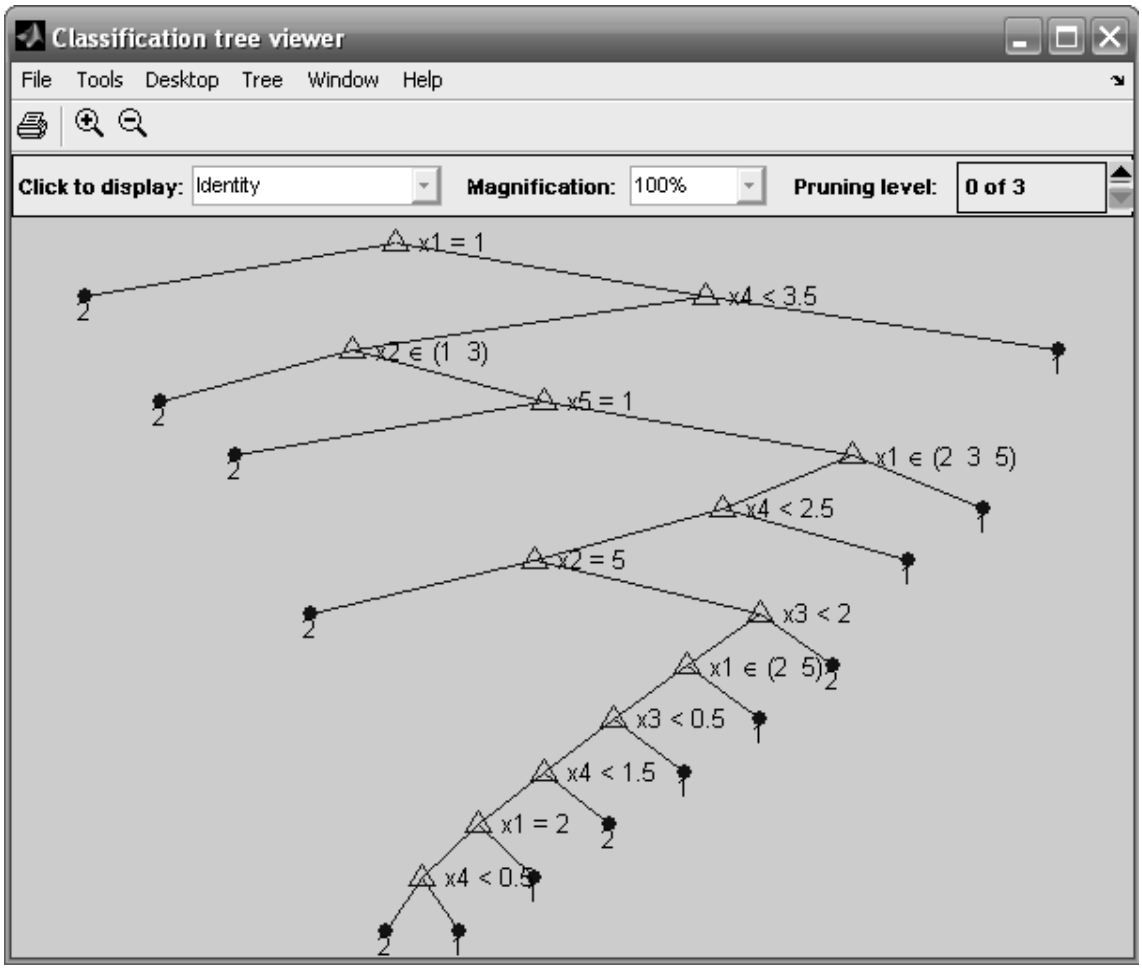

Рис. 1. Дерево рішень на основі повної вибірки

Проте для повноти експерименту нам потрібно розділити нашу вибірку на тренувальну і тестову. Для тренувальної вибірки візьмемо всі непарні дописи, для тестової – парні. Цього разу будуємо дерево рішень на основі тренувальної вибірки. В цьому випадку дерево рішень також використовує всі запропоновані параметри. Можна зробити висновок, що запропоновані вище параметри інформативні для визначення корисності допису.

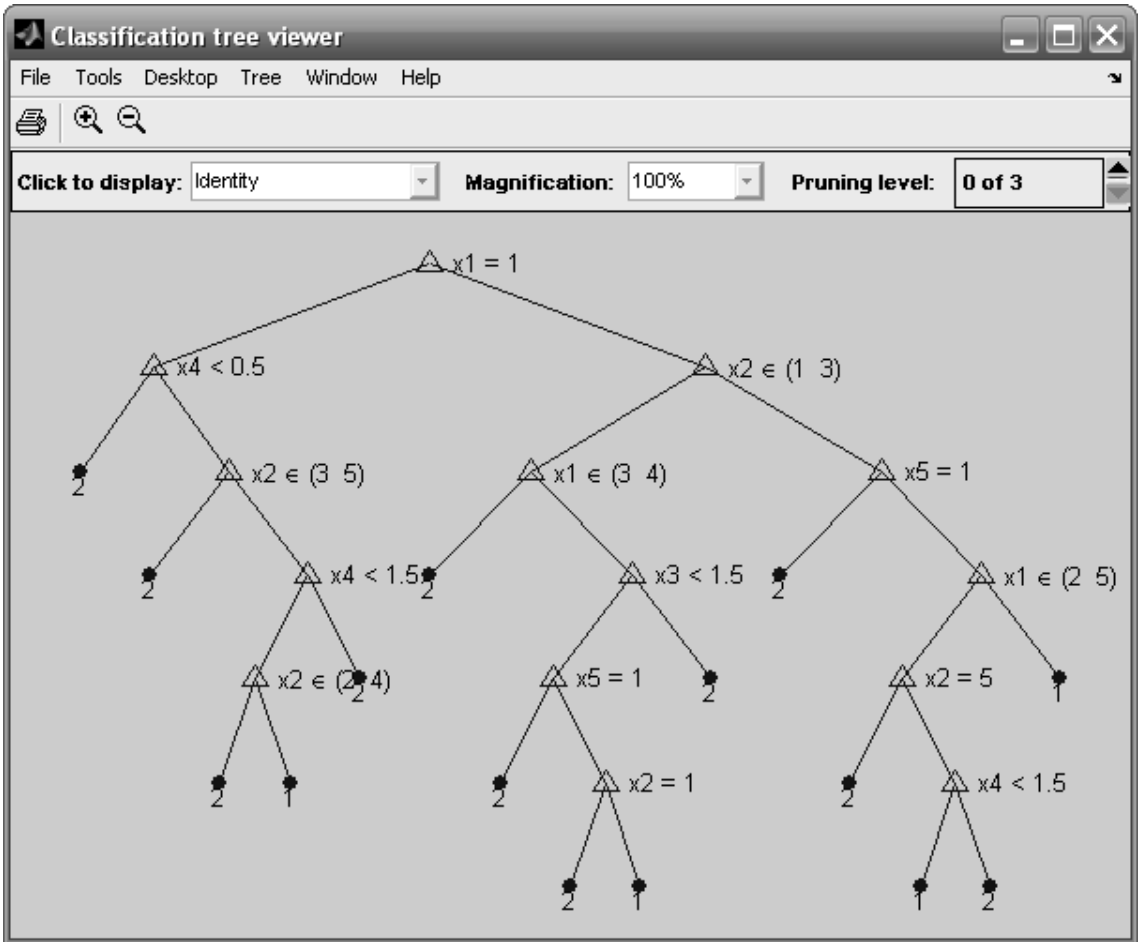

Рис. 2. Дерево рішень на основі тренувальної вибірки

Порівняємо результати, отримані на основі дерева рішень, а також дійсні дані. З таблиці результатів (табл. 3) бачимо, що правильно визначили 10 з 13 корисних дописів, а також 78 з 92 некорисних дописів. Отже, цей метод дає 83,8 % правильних відповідей.

Таблиця 3

Результати, отримані на основі дерева рішень

Чи корисний допис?	Так	Ні
Так	10	3
Ні	14	78

У майбутньому планується провести масштабніші експерименти з вхідними параметрами. Наприклад, використати додаткові параметри, що розкривають та деталізують базові або змінити формальне подання вибраних параметрів. Також потрібно проаналізувати більшу кількість різних дописів з різних предметних тематик сайта http://stackoverflow.com/_або аналогічних ресурсів, що технічно не викликає труднощів.

Висновки і перспективи подальших наукових розвідок

Перевагою запропонованого методу, на думку авторів, є його простота. На вході маємо html-код сторінки. Отже, вхідний текст не потрібно додатково розмічати. Для того, щоб розпарсити вхідний текст, також існують готові бібліотеки java, одна з яких використовується у цій роботі.

У майбутньому планується розглянути інші параметри, а також провести експерименти з уже запропонованими, щоб визначити, які з них найбільше впливають на якість допису.

1. The Quality of Content in Open Online Collaboration Platforms: Approaches to NLP-supported Information Quality Management in Wikipedia. Oliver Ferschkeю. July, 2014. 2. Automatic Scoring of Online Discussion Posts. Nayer Wanas, Motaz El-Saban, Heba Ashour, Waleed Ammar. WICOW '08 Proceedings of the 2nd ACM workshop on Information credibility on the web, pp. 19–26, 2008. 3. Automated essay scoring with e-rater v.2. Yigal Attali, Jill Burstein. The Journal of Technology, Learning, and Assessment, 4(3). February, 2006. 4. Automatically Assessing the Post Quality in Online Discussions on Software. Markus Weimer, Iryna Gurevych, Max Mühlhäuser In: Companion Volume of the 45rd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 125–128, 2007. 5. http://stackoverflow.com/ – Stack Overflow. 6. Точка, точка, запятая: машинное обучение. Хабрахабр. – http://habrahabr.ru/company/mailru/blog/112142/. 7. NLP: проверка правописания – взгляд изнутри (часть 2) – http://geektimes.ru/post/108923/. 8. Список функций Statistics Toolbox. Описание функции Treedisp. – http://matlab.exponenta.ru/statist/book2/17/treedisp.php. 9. Список функций Statistics Toolbox. Описание функции Treefit. – http://matlab.exponenta.ru/statist/book2/17/treefit.php. 10. Список функций Statistics Toolbox. Описание функции Treeprune. – http://matlab.exponenta.ru/statist/book2/17/treeprune.php.