КОМП'ЮТЕРНІ СИСТЕМИ ТА МЕРЕЖІ

Vol. 2, No. 1, 2020

https://doi.org/10.23939/csn2020.01.053

УДК 004.75:004.896

МЕТОДИ СЕМАНТИЧНОГО АНАЛІЗУ ПРИ АНОТОВАНОМУ УЗАГАЛЬНЕННІ ТЕКСТОВИХ ДОКУМЕНТІВ

О. Б. Очерклевич, А. О. Ігнатович

Національний університет "Львівська політехніка",
кафедра електронних обчислювальних машин

Розглянуто використання семантичного аналізу при узагальненні текстових документів. Проаналізовано найпоширеніші методи узагальнення текстових документів та оцінювання якості результатів оцінювання. Наведено особливості вдосконаленого методу анотаційного узагальнення текстових документів, який використовує принципи прихованого семантичного аналізу та елементи нечіткої логіки для виявлення семантично важливих речень. Запропоновано використання нового підходу до оцінювання ефективності узагальнення, основаного на елементах нечіткої логіки та на статистичному показнику, що використовується для оцінювання важливості слів у контексті та класу документа, що дає змогу визначити міру відповідності вмісту оригінального документа та його резюме. Наведено результати верифікації запропонованих засобів, що засвідчують їхню ефективність.

Ключові слова: текстовий документ, анотоване узагальнення, семантичний аналіз, нечітка логіка, оцінювання, ефективність

Вступ

Анотаційне узагальнення текстових документів займає значне місце в комп'ютерному опрацюванні текстової інформації. Фактичний величезний обсяг електронної інформації повинен бути зменшений, щоб користувачі могли ефективніше обробляти цю інформацію. Дослідження в царині анотаційного узагальнення текстових документів скеровані на вирішення двох основних завдань. Першим завданням є мінімізація обсягу текстового документа. Другим завданням є забезпечення максимальної відповідності вмісту між оригінальним документом та результатом анотаційного узагальнення. Це взаємовиключні завдання, які не мають оптимального вирішення. Сьогодні відомі окремі результати досліджень щодо методів як анотаційного узагальнення, так і методів оцінювання міри відповідності вмісту оригінального документа та його резюме [1–8]. Кожний з відомих методів має позитивні аспекти та недоліки. Відповідно важливим є аналіз позитивних аспектів та недоліків відомих методів, що дозволяє вибрати найкращий для конкретного застосування або розробити нові підходи чи методи.

Стан проблеми

Загальні підходи до анотування тексту з використанням семантичного аналізу можна поділити на чотири класи [1, 7, 8]. Перший клас – це евристичні підходи. Це методи вилучення, які використовують для підрахунку речень легкі прийоми. Наприклад, позицію речення в документі або входження слова із заголовка в речення . Наступна група включає підходи, основані на корпусі

документів (методи, основані на корпусі). Прикладом такого методу є TF-IDF (частота термінів – зворотна частота документа) [7]. Третій клас складається з методів, що враховують структуру дискурсу. Прикладом може бути метод лексичних ланцюгів, який здійснює пошук ланцюжків контекстних слів у тексті [8]. Четвертий клас містить підходи, багаті на знання. Конкретні методи анотування ґрунтуються на цих підходах.

При реалізації конкретного завдання окрім універсальних загальних підходів мають використовуватися спеціалізовані підходи до конкретних застосувань. Найпоширеніші спеціалізовані узагальнення текстових документів переважно основані на семантичному аналізі [1–8]. Застосовують нейронні мережі для опрацювання текстових документів [9, 10]. Відсутність загальноприйнятого оптимального методу анотування текстових документів є підставою для проведення подальших досліджень щодо ефективних методів та засобів вирішення цього наукового завдання.

Постановка задачі

Розглянути методи семантичного аналізу, узагальнюючи текстові документи та дослідити найефективніший для анотаційного узагальнення.

Результати досліджень

Ідею використання прихованого семантичного аналізу (Latent Semantic Analysis – LSA) для узагальнення тексту опубліковано у 2001 році [3, 8]. Застосована декомпозиція єдиного значення сингулярного розкладання матриці (Singular value decomposition – SVD) для узагальнення тексту. Процес починається із створення термінів матрицею речень $A = [A_1, A_2, ..., A_п]$. Кожний вектор стовпця A_i представляє зважений вектор терміно-частоти речення у розглянутому документі. Якщо в документі є загалом m термінів і n речень, тоді ми матимемо матрицю m × n для цього документа. Оскільки кожне слово зазвичай не зустрічається в кожному реченні, матриця A є розрідженою. Враховуючи розмірність m × n при m ≥ n для матриці A без втрати загальності, сингулярне розкладання матриці A визначається на підставі формули [3, 8]:

$$A=U\Sigma V^T,$$

де $U = [u_{ij}]$ – m × n ортонормальна матриця стовпців, стовпці якої називаються лівими сингулярними векторами;

Σ – діагностування $(\sigma_1, \sigma_2,\ldots \sigma_n)$, Σ_n є n × n діагональною матрицею, діагональні елементи якої є невід'ємними сингулярними значеннями, відсортованими за спаданням;

$V^T = [v_{ij}]$ є n × n ортонормальною матрицею, стовпці якої називаються правими сингулярними векторами.

Якщо ранг матриці (A) = r, то Σ задовольняє умову:

$$\sigma_1 \geq \sigma_2 \ldots \geq \sigma_p > \sigma_{p+1} = \ldots = \sigma_n = 0.$$

Структурну схему розкладання одиничного значення зображено на рис. 1.

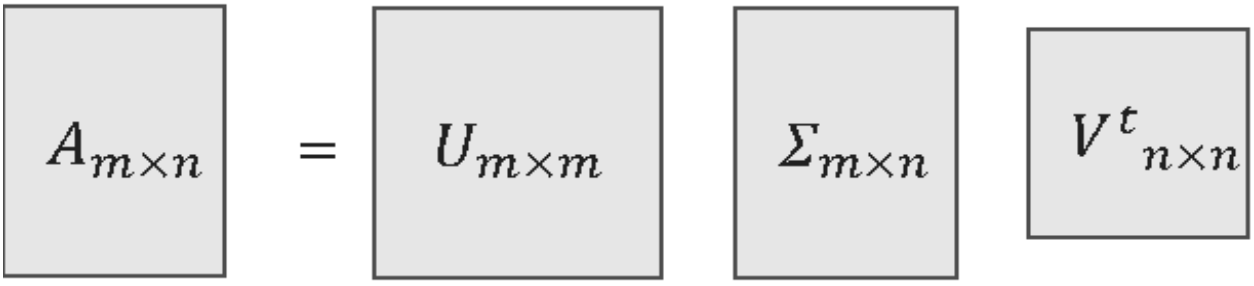

Рис 1. Структурна схема розкладання одиничного значення

Інтерпретувати застосування сингулярного розкладання матриці A до термінів речень можна за двома підходами. Стосовно трансформації текстового документа сингулярне розкладання матриці отримує відображення між m-мірним простором, породженим зваженими векторами частотних значень, і r-мірним сингулярним векторним простором. Стосовно семантичного аналізу сингулярне розкладання матриці формує приховану семантичну структуру з документа, представленого матрицею A. Ця операція відображає розбиття оригінального документа на r лінійно незалежних базових векторів або понять. Кожен термін і речення з документа спільно індексуються цими базовими векторами чи поняттями. Унікальна особливість сингулярного розкладання матриці полягає в тому, що воно здатне фіксувати та моделювати взаємозв'язки між термінами, забезпечуючи можливість семантично кластерувати терміни та речення. Окрім того, якщо шаблон поєднання слів є помітним і повторюється в документі, цей шаблон буде захоплений і представлений одним із набору векторів. Величина відповідного одиничного значення вказує на ступінь важливості цього шаблону в межах документа. Будь-які речення, що містять цей шаблон поєднання слів, проектуватимуться по цьому єдиному вектору, і речення, яке найкраще відображає цей шаблон, матиме найбільше значення індексу з цим вектором. Оскільки кожен конкретний шаблон поєднання слів описує певну тему чи концепцію в документі, описані вище факти, природно, призводять до гіпотези, згідно з якою кожен одиничний вектор представляє помітну тему концепцію документа, а величина відповідного йому одиничного значення представляє ступінь важливості конкретної теми чи концепції.

Розглянутий метод відносно має два недоліки, які суттєво впливають на його ефективність. Спочатку необхідно використовувати ту саму кількість вимірів, що і кількість речень, з яких потрібно вибрати невелику кількість для короткого викладу. Однак, чим більша кількість розмірів зменшеного простору, тим менш значущою темою ми беремо підсумок. Цей недолік перетворюється на перевагу лише у тому випадку, коли завчасно відомо, скільки різних тем має оригінальний документ. Відповідно у резюме підбирається однакова кількість речень. Другий недолік полягає в тому, що речення з великим значенням індексу, але не найбільшим (воно не отримає переваги в жодному вимірі), не буде обрано, хоча його зміст для резюме підходить одним із найкращих варіантів.

Для усунення обговорених недоліків запропоновано наступні модифікації методу узагалінення на основі сингулярного розкладання матриці. На першому етапі потрібно обчислити SVD термінів за матрицею речень. У результаті отримуємо три матриці, як показано на рис. 1. Для кожного вектора речень у матриці V^T (його компоненти помножуються на відповідні одиничні значення) обчислюємо його довжину. Причиною множення є надання переваги значенням індексу в матриці V^T, які відповідають найвищим одиничним значенням (найбільш значущим темам).

Другий модифікований метод узагальнення текстових документів оснований на використанні елементів нечіткої логіки. Для кожного речення формується вектор фіксованої розмірності. Множина значень векторів речень є основою використання методів нечіткої логіки. Правила нечіткої логіки базуються на представленнях у виді IF-THEN. Застосування нечітких правил відображають вектор речення у наступні значення: Low, Medium, High. Для прикладу розглянемо вектор довжини 5 елементів. Якщо перше значення довжини є High, друге значення вектора є High, третє значення вектора є Medium, четверте значення вектора є High та останнє значення є Medium, тоді речення має вагоме значення у формуванні теми у документі.

Відповідно, узагальнювач на базі нечіткої логіки кожному реченню формує оцінку важливості. Використовуючи цю оцінку, кожне речення впорядковується за спаданням відповідно до оцінки. Витягуються певна кількість речень з найбільшою оцінкою. Завершальний етап підсумовування тексту використовує речення з вибраними оцінками та розташовує речення у порядку, в якому вони були розташовані у оригінальному тексті.

Оцінювати важливість можна з використанням статистичного показника, який формує кількісне значення важливості речення у контексті та класі текстового документа. Прикладом застосування такого показника є відомий TF-IDF-CF метод [7]. Однак цей метод не враховує класи документів, що обмежує його застосування. Відповідно запропоновано використання нового параметра $\mathbf{a_{ij}}$ для представлення властивостей класу:

$$a_{ij}= \log(f_{ij}+1.0) *\log((N+1.0)/n_j) * n_{cij}/N_{ci} ,$$

де a_{ij} елемент матриці A; f_{ij} частота терміна у документах; n_j – число входжень терміну в документах; N – кількість документів колекції; n_{cij} – число входжень терміна у документи класу C; N_{ci} – число входжень терміна у всі класи.

Цей показник представляє частотну властивість класу, яка обчислює частоту терміна в документах одного класу.

Цей модифікований метод передбачає три основні етапи. Перший етап попередньої класифікації текстових документів відповідно до їх класів. Якщо клас докумета не відомий, тоді введемо для нього нову мітку класу, яка позначає невідомі класи. Після цього весь корпус текстової інформації представляється як вектор документа за допомогою методу TF-IDF-CF.

На другому етапі відбувається формування матриці A за допомогою TF-IDF-CF та застосування сингулярного розкладання матриці речень A. Ми отримуємо три матриці. Для кожного вектора речень у матриці V^T (його компоненти помножуються на відповідні одиничні значення) обчислюємо його довжину.

На третьому етапі, коли ветори метриці V^T є сформованими та представляють важливість речення у темі, застосовується узагальнювач на базі нечітої логіки.

Оцінювання якості семантичного аналізу при анотаційному узагальненні текстових документів є складним завданням. Це не менш важлива сфера порівняно із безпосереднім процесом узагальнення. Тому доцільно розглянути декілька найпоширеніших підходів для такого оцінювання [7].

Оцінюванняа за спільним відбором речень. Заходи спільного відбору містять точність та згадування вибраних речень. Ці методи вимагають наявності в розпорядженні оцінювача “правильного витягу”, для якого можна обчислити якість. Можна отримати цей екстракт кількома способами. Найпоширеніший спосіб – отримати деякі людські (ручні) екстракти та оголосити середнє значення цих екстрактів як “ідеальний (правильний) екстракт”. Однак отримання людських екстрактів зазвичай є проблематичним. Інша проблема полягає в тому, що два ручні зведення одного і того ж введення загалом не містять багато однакових речень.

Оцінювання на основі вмісту. Можна прояснити вищезазначену слабкість заходів спільного відбору за допомогою заходів схожості на основі вмісту. Ці методи обчислюють подібність між двома документами детальніше, ніж просто речення. Основним методом є обчислення подібності між повнотекстовим документом та його резюме з показником подібності косинусів, обчисленим за такою формулою [7]:

$$\cos(X,Y) = \frac{\sum x_i * y_i}{\sqrt{\sum (x_i)^2} * \sqrt{\sum (y_i)^2}},$$

де X та Y – подання на основі векторної просторової моделі.

Оцінювання на основі завдань. Встановлюють результати діяльності людини, створюючи та використовуючи конспекти для певного завдання. Наприклад, ми можемо виміряти придатність використання резюме замість повних текстів для категоризації тексту . Для цього оцінювання потрібен класифікований набір текстів.

Оцінювання за подібністю основної теми. Цей метод порівнює перші ліві одиничні вектори повнотекстового сингулярного розкладання матриці, виконаного на вихідному документі, та зведеного сингулярного розкладання матриці, виконаного на резюме. Ці вектори відповідають найбільш помітному шаблону слів у повному тексті та його резюме [7].

Оцінювання на основі прихованого семантичного аналізу. Цей новий метод можна класифікувати до категорії, що базується на вмісті, оскільки він оцінює якість резюме через схожість вмісту між повним текстом та його резюме. Наш метод використовує декомпозицію термінів за одиницею значення за матрицею речень U. Ця матриця відображає ступінь важливості термінів у відомих темах чи концепціях. В оцінці ми вимірюємо схожість між матрицею U, отриманою сингулярним розкладанням на вихідному документі, та матрицею U, отриманою сингулярним розкладання на резюме. Для оцінки подібності необхідни виконати два етапи.

Оцінювання за методом експертних даних. За цим методом декілька експертів встановлюють міру відповідності (експертні дані) між оригінальним документом та його анотацією на більш детальному рівні, ніж просто речення. Експертні дані формуються на підставі оцінювання результатів узагальнення. За результат приймається середнє значення експертних даних усіх експертів.

Верифікація запропонованих засобів здійснена на документах із колекції Вікіпедії. Середня довжина текстових документів становила 50 речень. Коефіцієнт узагальнення був заданий на рівні 20%. Використано найбільш ефективні оцінювання за подібністю основної теми, оцінювання на основі вмісту, оцінювання на основі прихованого семантичного аналізу, оцінювання за методом експертних даних. Найкращі показники отримано для узагальнувача на основі методу прихованого семантичного аналізу та використання елементів нечіткої логіки.

Висновки

Аналітичний огляд літературних джерел підтвердив доцільність досліджень у царині анотаційних узагальнень текстових документів. Дослідження переваг та недоліків широко вживаних методів узагальнення дозволило встановити, що сьогодні відсутні оптимальні засоби виконання таких робіт. Запропоновано використання узагальнувача на основі методу прихованого семантичного аналізу та використання елементів нечіткої логіки, що дає деякі переваги щодо якості узагальнення порівняно з іншими розглянутими методами. Дослідження методів оцінювання якості семантичного аналізу при анотаційному узагальненні текстових документів показало складність практичної реалізації таких завдань. Рекомендовано використання найбільш ефективних оцінювань за подібністю основної теми, на основі вмісту, на основі прихованого семантичного аналізу, за методом експертних даних.

Подальші дослідження доцільно спрямувати на методи анотаційного узагальнення для великих текстових документів та оцінювання ефективності таких засобів.

Список літератури

1. Ahmad K., Vrusias B. PCF Oliveira: Final evaluation and categorization of the text. Proceedings of the 26th Annual ACM SIGIR International Conference on Information Search Research and Development, Toronto, Canada, 2003. pp. 443–444.

2. Ginek J., Hedgehog K.: A practical approach to automatic generalization of the text. Proceedings of the ELPUB '03 Conference, Guimarães, Portugal, 2003, pp. 378–388.

3. Gong Yu., Liu X.: Generalization of the general text by means of measurement of relevance and the hidden semantic analysis. Proceedings of the 24th Annual ACM SIGIR International Conference on Information Research and Development, New Orleans, Louisiana, USA, 2001. pp. 19–25.

4. HP Edmundson: New methods for automatic removal. Journal of the Association of Computers 16 (2), 2001. pp 228–264.

5. Kupiek J., Pedersen J., Chen F.: Summary: Proceedings of the Eighteenth Annual ACM SIGIR International Conference on Information Search Research and Development, Seattle, Washington, USA, 1995. pp. 68–73

6. Radev R., Teufel S., Saggion H., Lam V., Blitzer J., Qi H., Celebi A., Liu D., Drabek E.: Problems of evaluation in a large generalization of documents. Issue 41st Annual Meeting of the Association of Computational Linguistics, Sapporo, Japan, 2003. pp. 375–382.

7. Understanding Inverse Document Frequency: On theoretical arguments for IDF. Stephen Robertson. Reprinted from: Journal of Documentation 60, No. 5, pp. 503–520.

8. Using Latent Semantic Analysis in Text Summarization and Summary Evaluation. Josef Steinberger, Karel Jezek/https://www.researchgate.net/publication/313673360.

9. Design System of Image Recognition Based on Neural Network / Vitaliy Yarkun, Yaroslav Paramud and Roman-Andriy Ivantsiv // 15th International Conference. The Experience of Designing and Application of CAD Systems (CADSM`2019). Polyana (Svalyava), Ukraine, February 26 – March, 2019. – pp. 2/41–2/44.

10. Paramud Y., Yarkun V. Method rozpiznavannya symvoliv na zobragennyakh na osnovi zhortkovoii neiironnoi meregi./ Transactions on Computer systems and networks, Lviv Polytechnic National University Press, 2018, No. 905, pp. 96–105 (in Ukrainian).

METHODS OF SEMANTIC ANALYSIS IN ANNOTATED GENERALIZATION OF TEXT DOCUMENTS

O. Ocherklevich, A. Ihnatovych

Lviv Polytechnic National University,
Computer Engineering Department

The article is devoted to the use of semantic analysis in the generalization of text documents. The analysis of features of the most widespread methods of generalization of text documents and an estimation of quality of results of an estimation is carried out. Features of the improved method of annotative generalization of text documents, which uses the principles of hidden semantic analysis and elements of fuzzy logic to identify semantically important sentences, are presented. It is proposed to use a new approach to evaluating the effectiveness of generalization, based on elements of fuzzy logic and a statistical indicator used to assess the importance of words in the context and class of the document, which allows to determine the correspondence between the original document and its summary. The results of verification of the proposed tools, certifying their effectiveness.

Keywords: text document, annotation generalization, semantic analysis, fuzzy logic, evaluation, efficiency.