УДК 004.5

Метод побудови дайджесту, що базується на дискримінантній вазі опорних слів

Ланде Д.В.

Постановка проблеми

Вибір найбільш інформаційних, релевантних інформаційній потребі користувача документів – задача неоднозначна. Обмежену сукупність таких документів, оформлену у зручному для сприйняття вигляді звичайно називають дайджестом. Зазвичай дайджести формують люди, експерти-аналітики. Автоматизація цього процесу здійснюється на засадах різноманітних лінгвостатистичних алгоритмів, серед яких відомі алгоритми кластерного аналізу, передусім K-means [1], LSA [2], алгоритми, що базуються на текстових маркерах (вони критично залежать від мови документів), мережевих [3], гібридних.

Загальна проблема всіх існуючих алгоритмів – великий час виконання, пов'язаний із такими проблемами: 1) обчислювальною складністю; 2) отриманням якісного дайджесту; 3) оцінюванням якості створеного дайджесту. Автором вже було запропоновано оцінювати якість реферування на основні теорії інформації (дивергенції Дженсена-Шеннона) [3], що частково розв'язує 3 проблему. Запропонований в цій роботі алгоритм за рахунок врахування семантичних маркерів текстів і практично лінійної обчислювальної складності дає частковий розв'язок проблеми 1) і 2).

Мета

Метою роботи було створення методу формування дайджесту на базі автоматичного аналізу масиву документів (повідомлень) із соціальних мереж, що базується на врахуванні дискримінантної ваги опорних слів із цих документів.

Обґрунтування

Передумовою для створення методу було те, що у розпорядженні у автора є документи, зібрані із соціальних мереж за допомогою системи «КіберАгрегатор» [4]. Кожному із цих документів поставлено у відповідність так званні «опорні слова», заздалегідь задана кількість найбільш вагомих слів із документів, визначених методом CHVG [5].

Передбачається, що за запитом до системи «КіберАгрегатор» формується відповідний інформаційний масив документів, з яких необхідно вибрати найвагоміші у деякому сенсі, з яких саме й має складатися дайджест.

Здавалося б, що самі найчастотніші слова із множини всіх опорних слів, що входять до релевантних документів, і мають бути визначними для відбору документів для дайджесту. Але ж існує ще одна вимога, а саме, дайджест має висвітлювати різні аспекти інформаційної потреби користувача, тобто окремі документи як складові дайджесту мають максимально змістовно відрізнятись. Ця властивість має враховуватись шляхом урахування входження в документи із дайджесту опорних слів з найбільшою дискримінантною вагою. Визначення цієї ваги опорного слова w_i може бути надано як добуток абсолютної частоти його появи tf_i і не спадної функції F від долі опорних слів x, з якими це слово не перетинається (не входить до однакових документів):

$$w_i = F(tf_i) \cdot (x+1).$$

Будемо вважати, що документи, що містять окремі слова з найбільшою дискримінантною вагою, а також мають найбільшу сумарну частотну вагу за іншими опорними словами, що входять до них, і складають основу дайджесту.

Алгоритм

Для побудови дайджесту застосовується алгоритм, що містить такі кроки:

1. Формується тематичний масив документів, що відповідають запиту користувача. Вважається, що кожному документу заздалегідь приписані опорні слова, найбільш вагомі слова з цього документа за алгоритмом CHVG.
2. Розраховується абсолютна частота появи кожного опорного слова, вибирається задана кількість найбільш вагомих, наприклад TOP-100.
3. Формується матриця взаємної появи в документах відібраних опорних слів. Елементом матриці є кількість взаємних поява пар слів. Вочевидь, діагональ цієї матриці відповідає абсолютній частоті цих слів.
4. На основі сформованої матриці і наведеної вище формули (як функція *F* обирається квадратний корінь) для кожного слова розраховується його дискримінантна вага.
5. Розраховується вага кожного із отриманих на кроці 1 документів як сума частотних вагових значень відповідних їм опорних слів.
6. Виводиться необхідна кількість найвагоміших документів, що містять опорні слова із найбільшою дискримінантної вагою. Ці документи складають дайджест.

Висновки

Алгоритм передбачає роботу із заздалегідь визначеними опорними словами, за рахунок чого забезпечується швидкість і якість. Цей алгоритм, як і відомий алгоритм LSA, є алгоритмом кластерного аналізу, в якому застосовується матричне представлення даних.

Новизна алгоритму полягає в визначенні формули для розрахунку дискримінантної ваги опорних слів, найвагоміші з яких по суті виступають центрами для визначення кластерів – центроїдами.

Складність наведеного алгоритму лінійна, він також може виступати як основа первинного вибору центроїдів для іншого швидкозбіжного алгоритму – K-means.

У відповідності із наведеним алгоритмом було розроблено інструментальні засоби формування дайджестів, які вбудовано в систему «КіберАгрегатор».

Література

1. K. P. Sinaga and M. Yang, "Unsupervised K-Means Clustering Algorithm," in IEEE Access, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

2. Salloum S.A., Khan R., Shaalan K. (2020) A Survey of Semantic Analysis Approaches. In: Hassanien AE., Azar A., Gaber T., Oliva D., Tolba F. (eds) Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020). AICV 2020. Advances in Intelligent Systems and Computing. – Vol 1153. Springer, Cham. https://doi.org/10.1007/978-3-030-44289-7_6

3. Lande D., Yang Zijiang, Zhu Shiwei, Guo Jianping, Wei Moji. Chinese legal information automatic summarization // CEUR Workshop Proceedings (ceur-ws.org). – Vol-2318/ Selected Papers of the XVIIIШ International Scientific and Practical Conference on Information Technologies and Security (ITS 2018). – P. 222-238

4. Lande D., Subach I., Puchkov A. System of Analysis of Big Data from Social Media // Information & Security: An International Journal 47, no. 1 (2020): 44-61.

5. Lande D.V., Snarskii A.A., Yagunova E.V., Pronoza E.V. The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text // 12th Mexican International Conference on Artificial Intelligence, 2013. – P. 209-215.