УДК 004.021

М.О. ГРАНІК, В.І. МЕСЮРА
Вінницький національний технічний університет

МЕТОД ВИЗНАЧЕННЯ СХОЖОСТІ НОВИННИХ ТЕКСТІВ НА ОСНОВІ СТАТИСТИЧНОЇ МІРИ "TERM FREQUENCY-INVERCE DOCUMENT FREQUENCY"

Метою роботи є розробка методу визначення схожості новинних текстів. У роботі запропоновано метод порівняння схожості новинних текстів на основі статистичної міри "term frequency – inverse document frequency", наведено результати його застосування. Метод може бути використано для кластеризації новинних текстів.

Ключові слова: новини, порівняння новин, tf-idf.

M.O. GRANIK, V.I. MESYURA
Vinnytsia National Technical University

METHOD OF EVALUATING THE SIMILARITY OF THE NEWS ARTICLES BASED ON STATISTICAL MEASURE "TERM FREQUENCY-INVERCE DOCUMENT FREQUENCY"

Abstract – The purpose of the paper is developing of the method of evaluating the similarity between news articles. This paper suggests the method of comparing the similarity of few news articles based on statistics measure term frequency – inverse document frequency". The results of the software, that implements suggested method, are also in the paper. The method may be used for clusterization of the news articles

Keywords: news, news comparing, tf-idf.

Вступ

Проблема визначення схожості новинних текстів є дуже актуальною проблемою. Отримання числової міри схожості новинних текстів може бути ефективно використана для задачі кластеризації новин. Кластеризація новин, у свою чергу, є важливою практичною проблемою, адже вона може бути використана у агрегаторах новин та у системах оцінювання правдоподібності новинної інформації.

Аналіз існуючих методів визначення схожості текстів довільної тематики

Сучасна наука пропонує декілька шляхів визначення числової міри схожості текстів.

Одним із таких шляхів є визначення косинусного коефіцієнта [1]. Для визначення цього коефіцієнта по кожному із текстів будується відповідний вектор, що містить інформацію про входження слів у кожен із цих текстів (наприклад, кількість цих входжень). Існує можливість порахувати скалярний добуток цих векторів двома способами (шляхом обрахунку суми добутку відповідних координат векторів, а також добутку довжин векторів на косинус кута між ними), і відповідним чином знайти косинус кута між ними. Так як описані вище вектори містять лише невід'ємні елементи, значення косинусу знаходиться у межах [0; 1]. Чим ближчим є отримане значення до 0, тим більш схожими є вектори і відповідні ним тексти.

Ще один шлях визначення схожості текстів – обрахування коефіцієнту Жаккара [2]. У відповідність кожному із текстів ставиться множина слів даного тексту. Коефіцієнт Жаккара визначається як частка від ділення потужності множини перетину двох отриманих множин на потужність множини об'єднання даних множин. Чим ближчим до одиниці є значення коефіцієнту Жаккара, тим більш схожими вважаються тексти.

Доволі схожим чином обраховується коефіцієнт Соренсена [2]. Аналогічно до методу Жакара, у відповідність кожному із текстів ставиться множина слів даного тексту. Значення коефіцієнту Соренсена рівне частці від ділення перетину двох отриманих множин на потужність мультимножини, що складається з двох даних множин (таким чином, при об'єднанні множин слів текстів кожне слово входить у мультимножину стільки раз, скільки воно зустрілось в обох множинах).

Також для обрахунку схожості двох текстів можуть бути застосовані коефіцієнт Сімпсона, коефіцієнт Браун-Бланке, коефіцієнт Кульчинського [2].

Недоліками перерахованих методів є те, що вони створювались для порівняння двох довільних текстів, в той час як тексти, що містять новинну інформацію, мають деякі суттєві особливості. Одна із найсуттєвіших особливостей (у порівнянні із, наприклад, художніми текстами) – часто заголовок однозначно визначає тему усього новинного тексту, адже це дозволяє привернути увагу читача до конкретної події, а також дозволяє читачам швидко орієнтуватись у великому обсязі новинної інформації. Метою роботи є розробка методу, що використовує таку важливу особливість.

Опис методу визначення схожості новинних текстів на основі статистичної міри tf-idf

Статистична міра *tf-idf*, на відміну від двох описаних вище методів, працює із набором (далі – корпусом) текстів [3]. Для того, щоб визначити, наскільки схожим є один із текстів на решту текстів із корпусу, для кожної пари «слово поточного тексту – текст, із яким відбувається порівняння» рахується частота входження слова у даний текст (term frequency, далі – *tf*). Також для кожного слова поточного тексту обраховується так звана зворотна частота документу (inverse document frequency, далі – *idf*). Формули для обрахунку цих двох величин виглядають наступним чином:

$$tf\left(t,d\right)=\frac{n_i}{\sum_k n_k} \quad (1)$$

$$idf\left(t,D\right)=log\frac{|D|}{|(d\supset t)|} \quad (2)$$

де t – поточне слово, d – поточний документ, n_i – кількість входжень поточного слова у поточний документ, D – корпус документів.

Як вже зазначалось, однією із найважливіших частин новинного тексту є його заголовок. Тому пропонується наступний метод визначення схожості деякого новинного тексту (далі – еталонного тексту) на інші тексти у даному корпусі:

1) Розглянути кожне слово із заголовку еталонного тексту.

2) Порахувати для кожного слова модифіковане значення $idf\left(t,D\right)$

3) Для кожного фіксованого слова розглянути всі документи корпусу, окрім еталонного. Розрахувати для кожної пари «слово-текст» модифіковане значення $tf\left(t,d\right)$.

4) До значення схожості даного тексту на еталонний додати величину $tf\left(t,d\right)$

Модифіковані значення $tf\left(t,d\right)$ та $idf\left(t,D\right)$ обраховуються із врахуванням важливості заголовку новин. Модифікована формула для обрахунку $tf\left(t,d\right)$ має наступний вигляд:

$$tf\left(t,d\right)=\frac{tn_i}{\sum_k n_k}, \quad (1),$$

де $t=c1\left(c1>1\right)$ якщо поточне слово входить до заголовку даного тексту; інакше $k=1$

Модифікована формула для обрахунку $idf\left(t,D\right)$ має наступний вигляд:

$$idf\left(t,D\right)=log\frac{\sum_{d_k\in D}q_i}{\sum_{d_k\in D,d_i\supset t}q_i}, \quad (2),$$

де $q_i=c2\ (c2>1)$ якщо поточне слово входить у заголовок даного тексту, інакше $q_i=1$. Застосування саме таких формул дозволяє надавати більшу вагу входженню слів до заголовку тексту. Регулювати цю вагу можна шляхом зміни коефіцієнтів $c1$ та $c2$. Питання знаходження оптимальних значень вищезазначених коефіцієнтів є нетривіальною задачею і може бути темою окремого дослідження. Перспективним напрямком у такому дослідженні можуть стати застосування інтелектуальних алгоритмів (генетичний алгоритм, алгоритм імітації відпалу, різноманітні ройові алгоритми тощо), а також перебору із відсіканнями.

Алгоритм визначення схожості новинних текстів на основі розробленого методу

На основі розробленого методу розроблено алгоритм визначення схожості новинних текстів, а також проведено його програмну реалізацію. Алгоритм має наступний вигляд:

1) Проведення операції стемінгу [4] над усіма текстами корпусу.

2) Вилучення із всіх текстів корпусу стоп-слів [5].

3) Обрахування для кожного тексту із корпусу (окрім еталонного) значення схожості даного тексту на еталонний. Для цього використовується описаний вище метод визначення схожості новинних текстів на основі статистичної міри *tf-idf*.

Стемінг – операція скорочення слів шляхом видалення із них неважливих частин, таких як префікс, суфікс чи закінчення (проте вважати, що в результаті застосування операції стемінгу кожне слово замінюється на його корінь, некоректно). Застосування алгоритмів стемінгу є поширеним у пошукових системах. Очевидно, що для порівняння текстів на схожість операція стемінгу також є надзвичайно важливою, адже вона дозволяє вважати різні форми одного і того ж самого слова (наприклад, слова у різних відмінках, числах тощо) одним і тим самим словом, що, у свою чергу, дозволяє отримати більш точну оцінку схожості текстів (в тому числі і новинних).

Стоп-слова (або шумові слова) – це такі слова у тексті, що не несуть змістовного навантаження. Під стоп словами зазвичай мають на увазі прийменники, частки, деякі інші окремі слова інших частин мови. Використання стоп-слів також часто застосовується у пошукових систем, однак їх використання є корисним для визначення схожості текстів. Так як стоп-слова не несуть змістовного навантаження, їх врахування при обрахунку схожості текстів можуть суттєво спотворювати отримані результати.

Для реалізації програмного продукту було використано такі значення коефіцієнтів:

$c1 = 1.2, c2 = 1.3$.

Для тестування було використано корпус із 30 новинних текстів (з урахуванням еталонного). 9 новин цього тексту освітлювали ту ж саму подію, що і еталонний, решта – довільну іншу тему. Графічне зображення отриманих результатів наведено на рисунку 1:

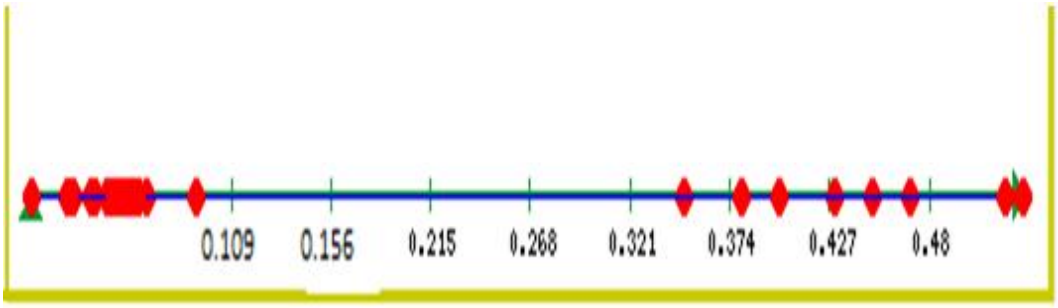

Рис. 1. Графічне зображення отриманих результатів

Можна побачити, що точки природнім чином розбились на два кластери, причому до складу кластеру, що знаходиться у правій частині малюнку, входять точки, що відповідають новинам, що описують ту ж саму подію, що й еталонна новина. Отже, можна зробити висновок, що розроблений алгоритм коректно оброблює вхідні тексти.

Висновки

Розроблено метод визначення схожості новинних текстів на основі статистичної міри *tf-idf*. Метод використовує важливість інформації, що подається у заголовку новинного тексту. На основі даного методу розроблено та реалізовано відповідний алгоритм. Отримані результати засвідчують коректність розробленого методу.

Розроблений метод може бути вдосконалено шляхом визначення оптимальних значень коефіцієнтів $c1$ та $c2$. Знаходження таких значень є складною задачею.

Література

1. Singhal Amit, Modern Information Retrieval: A Brief Overview, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4), P. 35–43.

2. Data Mining, University of Utah. URL: http://www.cs.utah.edu/~jeffp/teaching/cs5955/L4-Jaccard+Shingle.pdf

3. Karen Sparck Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, V 60, P. 493–502.

4. Lovins, Julie Beth, Development of a Stemming Algorithm, Mechanical Translation and Computational Linguistics 11, P. 22–31.

5. Text Mining, Analytics & More, All About Stop Words for Text Mining and Information Retrieval. URL: http://www.text-analytics101.com/2014/10/all-about-stop-words-for-text-mining.html

References

1. Singhal Amit, Modern Information Retrieval: A Brief Overview, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4), P. 35–43.

2. Data Mining, University of Utah. URL: http://www.cs.utah.edu/~jeffp/teaching/cs5955/L4-Jaccard+Shingle.pdf

3. Karen Sparck Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, V 60, P. 493–502.

4. Lovins, Julie Beth, Development of a Stemming Algorithm, Mechanical Translation and Computational Linguistics 11, P. 22–31.

5. Text Mining, Analytics & More, All About Stop Words for Text Mining and Information Retrieval. URL: http://www.text-analytics101.com/2014/10/all-about-stop-words-for-text-mining.html

Рецензія/Peer review : 8.5.2015 р. Надрукована/Printed :31.8.2015 р.
Рецензент: д.т.н., проф., Перевозніков С. І.