УДК 004.021

М. О. Гранік

МЕТОД ВИЗНАЧЕННЯ СХОЖОСТІ НОВИННИХ ТЕКСТІВ НА ОСНОВІ СТАТИСТИЧНОЇ МІРИ “TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY”

Вінницький національний технічний університет

Анотація

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: *слова: новини, порівняння новин, tf-idf*

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*

Опис методу визначення схожості новинних текстів на основі статистичної міри “term frequencey – inverse document frequency”

Статистична міра tf-idf працює із набором (далі – корпусом) текстів [1]. 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:

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

$$idf(t,D)=\frac{|D|}{|t\subset d|} \tag{2}$$

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.
3) Для кожного фіксованого слова розглянути всі документи корпусу, окрім еталонного. Розрахувати для кожної пари «слово-текст» модифіковане значення tf.
4) До значення схожості даного тексту на еталонний додати величину tf*idf.

Модифіковані значення tf та idf обраховуються із врахуванням важливості заголовку новин. Модифікована формула для обрахунку tf має наступний вигляд:

$$tf(t,d)=\frac{t\,n_i}{\sum_k n_k} \quad (3),$$

де $t=c1\,(c1>1)$ якщо поточне слово входить до заголовку данного тексту; інакше $k=1$.

Модифікована формула для обрахунку idf має наступний вигляд:

$$idf(t,D)=\log\frac{\sum_{d_i\in D} q_i}{\sum_{d_i\in D, d_i\supset t} q_i} \quad (4),$$

де $q_i=c2\,(c2>1)$ якщо поточне слово входить у заголовок данного 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 [2] над усіма текстами корпусу.

2) Вилучення із всіх текстів корпусу стоп-слів [3].

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, довзоляє 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 [4].

Для реалізації програмного продукту було використано такі значення коефіцієнтів: $c_1 = 1.25$, $c_2 = 1.3$.

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:

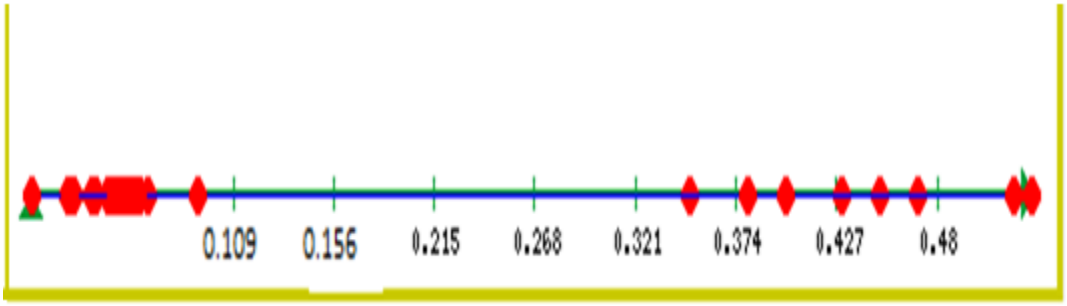

Рис. 1. Графічне зображення отриманих результатів

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 обробляє 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 c_1 та c_2. Знаходження таких значень є складною задачею.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. A statistical interpretation of term specificity and its application in retrieval / Karen Spärck Jones .— Journal of Documentation, V 60 .— P. 493-502.

2. Development of a Stemming Algorithm / Lovins, Julie Beth .— Mechanical Translation and Computational Linguistics 11 .— P. 22-31

3. Text Mining, Analytics & More. Стаття All About Stop Words for Text Mining and Information Retrieval. [Електронний ресурс] / Режим досутпу до статті: http://www.text-analytics101.com/2014/10/all-about-stop-words-for-text-mining.html

4. Метод визначення схожості новинних текстів на основі статистичної міри "Term frequency – Inverse document frequency" / М. Гранік, В. Месюра .— Вісник Хмельницького національного університету 4.2015 .— Ст. 180-182.

Михайло Олександрович Гранік — аспірант кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця, e-mail: Fcdkbear@gmail.com.

Науковий керівник: **Володимир Іванович Месюра —** кандидат технічних наук, доцент, професор кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця.

Mykhailo O. Granik — Postgraduate student of the Computer Science Chair, Vinnytsia National Technical University, Vinnytsia, e-mail: Fcdkbear@gmail.com.

Supervisor: **Volodymyr I. Mesyura —** Cand. Sc., Assistant professor, professor of the Computer Science Chair, Vinnytsia National Technical University, Vinnytsia.