УДК 004.021

М. О. Гранік
В. І. Месюра

МЕТОД ВИЗНАЧЕННЯ СХОЖОСТІ НОВИННИХ ТЕКСТІВ ШЛЯХОМ ПОРІВНЯННЯ ЇХ ЗАГОЛОВКІВ ІЗ ВИКОРИСТАННЯМ ЗАДАЧІ ПРО ПРИЗНАЧЕННЯ

Вінницький національний технічний університет;

Анотація

Метою роботи є розробка методу визначення схожості новинних текстів. У роботі запропоновано метод порівняння схожості новинних текстів на основі порівняння їх заголовків. Ця задача була зведена до задачі порівняння коротких текстів. Метод може бути використано для кластеризації новинних текстів.

Ключові слова: новини, порівняння новин, задача про призначення .

Abstract

The main goal of the article is the development of the method for comparing news articles. It is suggested to compare news articles based on their titles. This problem was reduced to assignment problem. Method can be used for news articles clasterization.

Keywords: news articles, comparison of the news articles, assignemnt problem.

Вступ

Проблема визначення схожості новинних текстів є дуже актуальною. Отримання числової міри схожості новинних текстів може бути ефективно використане для задачі кластеризації новин. Кластеризація новин, у свою чергу, є важливою практичною проблемою, адже результати її розв'язання можуть бути використані у агрегаторах новин та у системах оцінювання правдоподібності новинної інформації.

Часто основний зміст новини виражається її заголовком. Журналісти намагаються передати заголовком, про що саме йдеться у новинній статті для того, щоб зацікавити читачів. Виникає ідея порівнювати новинні тексти шляхом порівняння їхніх заголовків.

Заголовки новинних текстів зазвичай є доволі короткими. Частіше за все вони складаються з одного речення. Тобто у випадку порівняння заголовків новинних текстів ми маємо справу із порівнянням коротких текстів.

Проблема порівняння коротких текстів не є новою. Наприклад, у своїй роботі її розглядають Courtney Corley і Rada Mihalcea. Їхній підхід базується на жадібному групуванні пар слів текстів, що розглядаються [1]. Mihai Lintean та Vasile Rus у своїй роботі використовують іншу жадібну евристику, і таким чином покращують результати, отримані Courtney Corley та Rada Mihalcea.

Також автори цієї статті запропонували алгоритм порівняння схожості новин на основі поліноміального хешування [2].

Як відомо, використання жадібних евристик не завжди приодить до найкращих результатів. Мета цієї статті – розробити метод, що позбавлений цього недоліку. У статті показано спосіб зведення задачі про схожість коротких текстів до добре відомої задачі - задачі про призначення.

Знаходження семантичної схожості слів

Велика кількість методів , що знаходять семантичну схожість слів, базуються на базах знань, що складені людиною-експертом.

Одна з найбільш відомих таких баз для слів англійської мови – лексична база даних WordNet. У WordNet слова згруповані у синонімічні множини, що називаються синсетами (synsets), кожна з яких описує якесь значення чи концепцію. Синсети пов'язані між собою за допомогою лексико-

семантичних зв'язків, таких, як гіперонімія (аналог зв'язку IS-A, що використовується у методах штучного інтелекту).

Існує велика кількість метрик, що слугують для визначення семантичної схожості і використовують структуру бази WordNet (метрика Wu-Palmer, метрика Jiang та Conrath тощо) [1].

Задача про призначення

Задачу про призначення зазичай формулюють на основі наступного прикладу. Є N постачальниів деякого товару, а також N клієнтів, що хотять цей товар придбати. Для кожної пари "постачальник-клієнт" відома ціна, яку потрібно заплатити для того, щоб цей постачальник доставив товар цьому клієнтові. Кожен постачальник може доставити товар не більше ніж одному клієнтові. Потрібно таким чином згрупувати постачальників та клієнтів, щоб усі клієнти отримали товар, і сумарна ціна всіх використаних операцій була мінімальна (чи максимальна). Як бачимо, у класичному варіанті задача формулюється для випадку, коли клієнтів та постачальників однакова кількість. Така задача називається лінійною задачею про призначення. Але від задачі для N клієнтів та N постачальників легко перейти до задачі із N клієнтами та M постачальниками (шляхом додавання додаткових фіктивних ребер із нульовою вагою) [3].

Метод визначення схожості новинних текстів шляхом порівняння їх заголовків

Розглянемо два коротких фрагменти тексту, що є заголовками новинних текстів. Мета розробленого методу – навчитись знаходити числове значення схожості між цими заголовками. Також слід дослідити, яким є значення схожості для заголовків новин, що розповідають про одну і ту саму подію, а також для заголовків новин, що розповідають про різні події.

Для розв'язання поставленої задачі пропонується метод, що складається з наступних кроків:

1) Видалення стоп-слів із обох заголовків.

2) Кожному із двох текстів ставиться у вдповідність одна із доль дводольного графа.

3) Для кожної пари "слово у першому реченні-слово у другому реченні" знаходиться значення семантичної схожості між ними за допомогою однієї із описаних у розділі "Знаходження семантичної схожості слів". Між відповідними цим словам вершинами проводиться ребро із вагою, рівною значенню схожості.

4) Для отриманого дводольного графа знаходиться максимальне парування максимальної ваги, тобто розв'язується задача про призначення.

5) Отримане максимальне значення ціни нормалізується шляхом ділення отриманого результату на розмір меншої долі графу.

6) Отримане нормалізоване значення вважається значенням семантичної схожості між заголовками новинних текстів.

Побудова, реалізація та тестування алгоритму визначення схожості новинних текстів шляхом порівняння їх заголовків

На основі описанного методу було створено та реалізовано відповідний алгоритм.

Було проведено попарні порівняння між десятьма англомовними загловками новинних текстів – усього сорок п'ять порівнянь. П'ять з цих заголовків були заголовками до новин, що розповідають про одну й ту саму подію, решта – п'ять заголовків новин, що розповідають про інші різні події.

Для визначення семантичної схожості слів використовувалась метрика Wu-Palmer.

Для розв'язання задачі про призначення було використано метод із використанням максимального потоку максимальної вартості.

Результати тестування показали, що для заголовків новинних текстів, що розповідають про одну й ту саму подію, значення схожості за описанним вище методом було в середньому рівне 0.726. Для пар текстів, в яких йдеться про різні події, середнє значення схожості їх заголовків було рівне 0.285.

Таким чином, можна побачити, що різниця між цими числами є доволі суттєвою.

Висновки

Розроблено метод визначення схожості новинних текстів шляхом порівняння їх заголовків із використанням задачі про призначення. На відміну від методів, описаних Courtney Corley і Rada Mihalcea, а також Mihai Lintean і Vasile Rus, цей метод не використовує жадібних евристик. Групування слів із різних заголовків проводиться не жадібним чином, а базуючись на загальному правилі оптимальності. Досягти цього допомогло формулювання задачі про схожість коротких текстів у термінах задачі про призначення.

На основі даного методу розроблено та реалізовано відповідний алгоритм. Отримані результати засвідчують коректність розробленого методу.

Розроблений метод може бути вдосконалено шляхом використання інших метрик для визначення семантичної схожості між словами.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Mihai Lintean. Measuring semantic similarity in short texts through greedy pairing and word semantics / Mihai Lintean, Vasile Rus // Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference .– 2012.– P. 244-249
2. Гранік Михайло.Алгоритм визначення схожості текстів новин на основі поліноміального хешування / Михайло Гранік, Володимир Месюра // Вісник Вінницького політехнічного інституту .– 2016.– Ст. 75-79
3. Угорський алгоритм розв'язку задачі про призначення [Електронний ресурс] .– Режим досутпу до статті: http://e-maxx.ru/algo/assignment_hungary

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

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

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

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