МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ТА ОБЧИСЛЮВАЛЬНІ МЕТОДИ

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, м. Вінниця

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 (а саме – до задачі визначення їх еквівалентності). У свою чергу, ця задача була зведена до задачі про призначення – класичної задачі із області комп'ютерних наук, що може бути розв'язана угорським алгоритмом чи за допомогою знаходження максимального потоку мінімальної вартості. Метод може бути використано для кластеризації новинних текстів, у сервісах агрегації новинних текстів.

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

Аннотация. Целью работы является разработка метода определения сходства новостных текстов. В работе предложен метод сравнения сходства новостных текстов на основе сравнения их заголовков. Эта задача была сведена к задаче сравнения коротких текстов (а именно – к задаче определения их эквивалентности). В свою очередь, эта задача была сведена к задаче о назначениях – классической задачи из области компьютерных наук, которая может быть решена при помощи венгерского алгоритма или посредством нахождения максимального потока минимальной стоимости. Метод может быть использован для кластеризации новостных текстов, в сервисах агрегации новостных текстов.

Ключевые слова: новости, сравнение новостей, задача о назначении.

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 the problem of comparing of the short texts (namely, to the equivalency detection problem). This problem was reduced to an assignment problem – classical computer science problem, that can be solved with Hungarian algorithm or using the algorithms, that find minimum cost maximum flow. The method can be used for news articles clasterization and for news aggregators.

Key words: news articles, comparison of the news articles, assignemnt problem.

Вступ

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. При порівнянні новинних текстів виділяють два підвиди задач – визначення відношення еквівалентності (paraphrase relation) та виначення відношення слідування (entailment relation), коли один текст є логічним висновком із іншого. У цій статті буде розглянуто тільки відношення еквівалентності, адже легко бачити, що саме сенсова еквівалентність заголовків означає, що в самих текстах йдеться про одну й ту ж саму подію.

Актуальність

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.

Мета

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.

Як відомо, використання жадібних евристик не завжди прводить 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.

Задачі

1. Розробка метода порівняння новинних текстів, що не базується на жадібних алгоритмах.
2. Перевірка коректності створеного методу.

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].

Перевага такого методу очевидна – використовується джерело знань, складене людиною, що підвищує ступінь довіри до нього. Недоліком такого методу є те, що одне й те ж саме слово може мати декілька значень, і тому незрозуміло, до якого саме синсету його віднести. Проблема визначення правильного значення слова у тексті досі є складною задачею у задачах обробки природної мови. Однак цієї проблеми можна позбутись, якщо брати до уваги тільки ті значення слів, що максимізують отриману схожість між ними.

Задача про призначення

Задача про призначення – добре відома оптимізаційна задача. Вона належить до класу задач комбінаторної оптимізації.

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) [2].

Існує і більш формальне та лаконічне формулювання задачі. Дано повний двудольний зважений граф. Необхідно знайти у ньому максимальне парування мінімальної (максимальної) ваги (min/max cost matching) [3].

Ще одне формулювання є таким. Дана матриця a, що має n рядків та n стовбців. Потрібно знайти таку перестановку стовбців p, що мінімізує (максимізує) сумму a[i][p[i]] для всіх i від 1 до n [2].

Таку задачу можна розв'язувати декількома способами. Перший із них використовує угорський алгоритм, що був розроблений Гарольдом Куном у 1955р. Саме Кун дав цьому алгоритмові назву "Угорський" [2].

Ще один спосіб розв'язання цієї задаі використовує максимальний потік мінімальної вартості. Зручніше за все його пояснити у термінах дводольного графа. Кожному ребру надамо пропускну здатність, рівну одиниці, а також орієнтуємо їх у напрямку від першої долі до другої. Створимо фіктивну вершину – витік, з якої проведемо орієнтовані ребра у всі вершини першої долі із нульовою вагою та одиничною пропускною здатністю. Також створимо фіктивну вершину – стік, у яку проведемо орієнтовані ребра із нульовою вагою та одиничною пропускною здатність від усіх вершин другої долі. У такому графі знайдемо максимальний потік мінімальної (максимальної) вартості. Легко бачити, що так як ми знаходимо максимальний потік, то це забезпечить максимальну завантаженість вершин графу. (тобто, виражаючись термінами першого формулювання, наш алгоритм максимізує кількість клієнтів, що обслуговувались). Так як ми знаходимо потік мінімальної (максимальної) вартості, то можна побачити, що саме знайдена вартість і буде тією оптимальною ціною, що потрібно знайти для розв'язання задачі про призначення [3].

Ці методи мають свої переваги та недоліки. Угорський алгоритм є швидким (має обчсилювальну складність $O(N^3)$), проте є більш складним у реалізації. Алгоритм із використанням максимального потоку мінімальної вартості є повільнішим за угорський алгоритм (має обчислювальну складність $O(N^4)$), проте є більш простим у реалізації [2, 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.

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. Так як стоп-слова не несуть змістовного навантаження, їх врахування при обрахунку схожості текстів можуть суттєво спотворювати отримані результати [4].

Кожному із двох текстів ставиться у вдповідність одна із доль дводольного графа. Пи цьому, кожному із слів, що залишились після видалення стоп-слів, ставиться у відповідність одна вершина графа (якщо слово повторюється декілька разів – йому відповідатиме декілька вершин).

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.

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

Отримане максимальне значення ціни нормалізується шляхом ділення отриманого результату на розмір меншої долі графу (тобто на розмір максимального парування; так як граф є повним дводольним графом, то ці дві величини співпадають).

Формально кажучи, нормалізоване значення обраховується наступним чином:

$$S = \frac{MatchingWeight}{MatchingSize} \tag{1}$$

Нормалізація потрібна для того, щоб була можливість порівнювати між собою знайдені результати схожості для різних пар текстів. Без цього кроку неможливо переконатись, що метод працює для визначення схожості новинних текстів.

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 (схожість заголовків новинних текстів, що розповідають про різні події, складає приблизно тридцять дев'ять відсотків від схожості новинних текстів, що розповідають про одну й ту саму подію). Отже, розроблений метод дозволяє з високою ймовірністю ідентифікувати, чи розповідають новинні тексти про одну й ту саму подію.

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 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. Угорський алгоритм розв'язку задачі про призначення [Електронний ресурс] .– Режим досутпу до статті: http://e-maxx.ru/algo/assignment_hungary

3. Задача про призначення. Розв'язок за допомогою min-cost-flow [Електронний ресурс] .– Режим досутпу до статті: http://e-maxx.ru/algo/assignment_mincostflow

4. 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

Стаття надійшла: 07.06.2016.

Відомості про авторів

Гранік Михайло Олександрович – аспірант напряму “Інформаційні технології”, Вінницький національний технічний університет, м. Вінниця.

Месюра Володимир Іванович – кандидат технічних наук, професор кафедри комп'ютерних наук Вінницького національоного технічного університету, м. Вінниця.