УДК 004.93:159.95

О. В. Бісікало
О. В. Яхимович

ВИЗНАЧЕННЯ КЛЮЧОВИХ СЛІВ З ТЕКСТУ ПОВІДОМЛЕНЬ МІКРОБЛОГІВ

Вінницький національний технічний університет

Анотація

Запропоновано використання методу визначення ключових слів англомовного тексту на основі інструментальних засобів пакету DKPro Core для визначення ключових слів з повідомлень мікроблогів.

Ключові слова: метод, ключові слова, англійська мова, мікроблоги, лінгвістичний пакет, DKPro Core, синтаксичний аналіз.

Abstract

Use of the method of determining keywords for english texts based on DKPro Core package to determine keywords in microblogging messages.

Keywords: method, keywords, English, microblogs, linguistic package, DKPro Core, syntactic analysis.

Вступ

Як відомо, не всі слова в тексті рівнозначні. Є слова, які дозволяють представити текст у згорнутому вигляді, при цьому таке подання досить точно відображає зміст вихідного тексту.

Основний зміст документа (тексту) може бути виражений за допомогою певних слів, узятих безпосередньо з цього тексту. Як правило, до кожного розгорнутого тексту можна скласти цілий набір ключових слів різного обсягу (від 5 до 15 слів). Але взагалі кількість ключових слів може варіюватися в широких межах [1].

Метою роботи є застосування методу визначення ключових слів англомовного тексту на основі інструментальних засобів пакету DKPro Core для визначення ключових слів з повідомлень мікроблогів.

Результати дослідження

На сьогоднішній день однією з найважливіших і помітних областей Web, ключовим принципом якої є участь користувачів в роботі сайтів, є мережеві щоденники, або веб-логи, скорочено назва – блоги. Концептуальним розвитком блогів, обумовленим їх широкою соціалізацією, є мікроблоги, які мають ряд характерних особливостей: обмежена довжина повідомлень, велика частота публікацій, різноманітна тематика, різні шляхи доставки повідомлень і т.д.

Перший і найбільш відомий сервіс мікроблогів Twitter був запущений в жовтні 2006 року компанією Obvious з Сан-Франциско. Очевидно, що автоматизоване виділення найбільш значущих термінів (слів) з потоку повідомлень, що генерується співтовариством Twitter, має практичне значення як для визначення інтересів різних груп користувачів, так і для побудови індивідуального профілю кожного з них.

Однак треба зазначити, що класичні статистичні методи екстракції ключових термінів, засновані на аналізі колекцій документів, малоефективні в даному випадку. Це обумовлено надзвичайно малою довжиною повідомлень (до 140 символів), їх різноманітною тематикою і відсутністю логічного зв'язку між собою, а також великою кількістю рідко використовуваних абревіатур, скорочень та елементів специфічного мікросинтаксису.

Оскільки в Twitter не існує простого і зручного способу для групування «твітів» різних користувачів за тематикою, співтовариство користувачів застосовує власне рішення – хештеги, які схожі на інші приклади використання тегів (наприклад, для анотування записів у звичайних блогах) і дозволяють додати «твіти» в якусь категорію.

Хештеги починаються з символу «#», за яким слідує будь-яке поєднання дозволених в Twitter

символів без пробілів; найчастіше це слова або фрази, в яких перша буква кожного слова наведена з великої літери. Вони можуть зустрічатися в будь-якій частині «твіта», часто користувачі просто додають символ «#» перед будь-яким словом. При додаванні в повідомлення хештега воно буде відображатися при пошуку в потоці повідомлень Twitter за цим хештегом.

До неофіційних, але загальноприйнятих правил використання хештегів відноситься вибір для них термінів, релевантних темі повідомлення, а також додавання лише невеликої кількості їх в одне повідомлення. Все це дозволяє розглядати їх в якості потенційних термінів, які з достатнім ступенем імовірності відображають загальну тематику повідомлення.

Одним із завдань добування інформації з тексту є виділення ключових термінів, що з певною мірою достовірності відображають тематичну спрямованість документа. Автоматичне вилучення ключових термінів можна визначити як автоматичне визначення важливих тематичних термінів у документі. Це є однією з підзадач більш спільного завдання – автоматичної генерації ключових термінів, для якої отримані ключові терміни не обов'язково повинні бути присутніми в даному документі [2].

В останні роки було запропоновано багато підходів, які дозволяють проводити аналіз наборів документів різного розміру і витягувати ключові терміни, що складаються з одного, двох і більше слів.

Найважливішим етапом витягання ключових термінів є розрахунок їх ваг в аналізованому документі, що дозволяє оцінити їх значущість відносно один одного в даному контексті. Для вирішення цього завдання існує ряд методів, які умовно діляться на 2 групи: вимагають навчання (з вчителем) і не потребують навчання. Під навчанням мається на увазі необхідність попередньої обробки вихідного корпусу текстів з метою вилучення інформації про частоту знаходження термінів у всьому корпусі. Іншими словами, для визначення значущості терміна у цьому документі необхідно спочатку проаналізувати всю колекцію документів, до якої він належить. Альтернативним підходом є використання лінгвістичних онтологій, які є більш-менш наближеними моделями існуючого набору слів заданої мови. На базі обох підходів були створені системи для автоматичної екстракції ключових термінів, однак у цьому напрямку постійно ведуться дослідження з метою підвищення точності і повноти результатів, а також з метою застосування методів вилучення інформації з тексту для розв’язання нових задач [3].

Краща якість обробки тексту досягається лінгвістичними методами або ж при їх комбінації зі статистичними, тому систему автоматичного визначення ключових фраз з тексту природною мовою слід розробляти з використанням морфологічного словника (лексикону) і синтаксичних правил.

Результати парсерингу природних мов за допомогою сучасних лінгвістичних пакетів дозволяють на доступному програмному рівні оперувати синтаксичними зв’язками між словами окремого речення [4]. Одним з таких пакетів є DKPro Core – це набір програмних компонентів для обробки природної мови, що базується на Apache UIMA framework. Він був побудований з метою підвищення продуктивності дослідників, які працюють з автоматичним аналізом мови. Підхід DKPro Core полягає в тому, що дослідники повинні мати можливість зосередитися на своїх реальних наукових питаннях, а не на розробці технологій [5].

Враховуючи можливості лінгвістичного пакету DKPro Core пропонується застосувати його для визначення ключових слів англомовного тексту з повідомлень мікроблогів. При цьому пошук та використання хештегів, що після вилучення символу «#» перетворюються у абсолютній більшості на звичайні слова, можуть служити основою для побудови методів з вчителем, тобто з’являється можливість порівнювати отримані результати пошуку ключових слів з множиною хештегів мікроблогу. Оскільки лексична онтологія з множини хештегів фактично задається користувачами, перевага запропонованого підходу полягає у прискорені процесу знаходження ключових слів за рахунок виключення процедур попередньої обробки вихідного корпусу текстів.

Висновки

Застосовуючи підхід до визначення ключових слів, що базується на використанні додаткової інформації про складні залежності між членами англомовного речення, можна знаходити ключові слова в тексті повідомлень мікроблогів і порівнювати їх з хештегами заданими автором повідомлень. Такий підхід дозволить прискорити процес знаходження ключових слів методом з вчителем за

рахунок виключення процедур попередньої обробки вихідного корпусу текстів. Для функціональної реалізації аналізатора обрано популярний лінгвістичний пакет DKPro Core.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Ершов Ю. С. Выделение ключевых слов в русскоязычных текстах / Ю. С. Ершов // Молодежный научно-технический вестник. – М.: ФГБОУ ВПО "МГТУ им. Н.Э. Баумана", 2014. – № ФС77-51038. – С. 70-79.

2. Turney P. Learning to extract keyphrases from text / Turney P. // Technical report, National Research Council, Institute for Informational Technology. – 1999. – Pp. 143-147.

3. Коршунов А. В. Извлечение ключевых терминов из сообщений микроблогов с помощью Википедии / Коршунов А. В. // Труды Института системного программирования РАН. – 2011. – №20. – С. 102-115.

4. Бісікало О.В. Концептуальна модель системи образного аналізу і синтезу природно-мовних конструкцій / О.В. Бісікало // Математичні машини і системи. – 2013. – № 2. – С. 184–187.

5. Natural Language Processing: Integration of Automatic and Manual Analysis [Електронний ресурс]. – Режим доступу: http://tuprints.ulb.tu-darmstadt.de/4151/1/rec-thesis-final.pdf – Назва з екрану.

Олег Володимирович Бісікало — доктор технічних наук, професор, декан факультету комп'ютерних систем і автоматики, Вінницький національний технічний університет, Вінниця.

Олександр Вікторович Яхимович — аспірант кафедри автоматики та інформаційно-вимірювальної техніки, факультет комп'ютерних систем і автоматики, Вінницький національний технічний університет, Вінниця, e-mail: yahimovich.olexandr@gmail.com.

Oleg V. Bisikalo — Doctor of Engineering, Professor, Dean of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia

Alexander V. Yahimovich — Department Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia, email: yahimovich.olexandr@gmail.com.