УДК 519.7:004.912

Т. Б. Шатовська, к. т. н., доц.;
І. В. Каменєва студ.

КОМБІНОВАНИЙ ІЄРАРХІЧНИЙ ПІДХІД ДО КЛАСТЕРИЗАЦІЇ ДОКУМЕНТІВ

Запропоновано інтегрований ієрархічний підхід до класифікації тексту, заснований на дендрограмі та k-середніх кластеризації. Цей підхід дозволяє представити новий інтегрований метод ієрархічної кластеризації, який може класифікувати дані без попереднього задання кількості класів, що дозволяє, структуровано зберігати документи у комп'ютері. Цей підхід оснований на двох методах, які відносяться до області text і data mining. Першим етапом є попереднє оброблення документів яке, скорочує час обчислення і отримання якісного результату. Другим етапом є використання векторної моделі, яка дозволяє чітко визначити важливість слів у документі. Використано ієрархічну кластеризацію, в яку входять два методи: дендрограми і k-середніх.

Метод дендрограми дозволяє заздалегідь визначити кількість кластерів (тек), метод k-середніх відносить документи до певних кластерів. Завершальним етапом є використання методу дендрограми для створення ієрархічної послідовності документів усередині кожного кластера (теки).

Вступ

У даний час класифікація тексту є актуальною науково-дослідницькою проблемою. Методи класифікації текстів застосовуються у фільтрації документів, розпізнаванні спаму, автоматичному анотуванні, знятті неоднозначності (автоматичні перекладачі), складанні Інтернет-каталогів, класифікації новин, розподілі реклами, персональних новинах. Класифікація документів застосовується в фільтрації електронної пошти, маршрутизації процесу відправки пошти, фільтрації спаму, контролі новин, вибірковому поширенні інформації для споживачів, автоматичній індексації наукових статей, автоматичної популяції ієрархічних категорій Web -ресурсів, ідентифікації документального жанру, дослідженні кодування.

З кожним роком збільшується обсяг доступних користувачеві масивів текстової інформації на робочому комп'ютері, що сприяє більшій актуалізації завдання пошуку необхідних користувачеві документів у таких масивах. Для виконання цього завдання часто застосовуються різні тематичні класифікатори, рубрикатори тощо, які дозволяють шукати (автоматично або вручну) документи у невеликій підмножині документальної бази відповідної тематики [1], яка цікавить користувача.

У статті розглядається метод автоматичної кластеризації, який здатний без допомоги користувача кластеризовувати хаотичне розташування файлів в структурований набір документів відносно тематики [2]. Ця система допоможе зберігати інформацію на комп'ютері в належному вигляді і не витрачати час на систематизацію документів за відповідними тематиками.

Кластерний ієрархічний підхід

Існує дуже велика кількість класифікаторів. Найчастіше використовуються методи Байєсовської класифікації, метод опорних векторів і багато інших. Ці методи мають істотний недолік — навчання з вчителем (supervised learning). У даному випадку класифікація визначає документи до однієї або більше зумовлених категорій. У класифікації тексту новий текстовий документ призначається в один з наявних наборів документів класу. У цій роботі використовується текстова кластеризація — навчання без вчителя (unsupervised learning). Вона застосовується для знаходження нових заздалегідь невідомих структур. [3]

Ієрархічна кластеризація — процес організації даних в деревовидну структуру, яка основана на подібності цих даних. Цей метод дуже потужний і корисний для аналізу великих наборів даних. Основна ідея — створити набір елементів у дереві. Дерево має багато гілок, якщо елементи подібні один до одного, до них приєднуються короткі гілки, і навпаки, якщо їх подібність зменшується,

тоді збільшуються гілки.

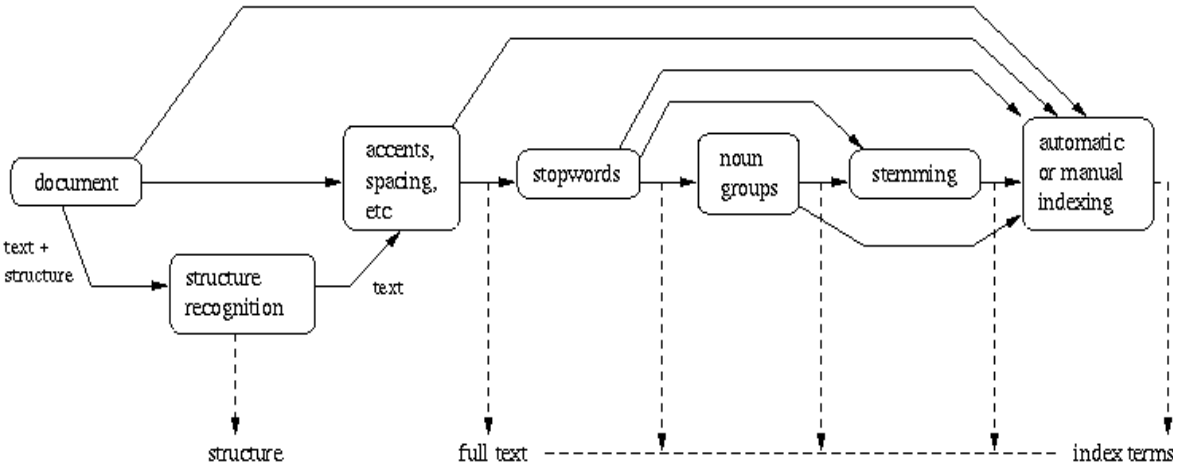
Припустимо, що існує декілька текстів. Необхідно згрупувати ці тексти відповідно до подібності їх стилів. Таке групування може бути як однорівневим («плоскою», з виділенням таких кластерів, що кожен об'єкт в них є одним з текстів набору кластеризації), так і ієрархічною, коли кластери, отримані в результаті об'єднання найбільш схожих текстів самі можуть об'єднуватися в кластери, а кластери кластерів — в інші кластери і так далі. Відношення тексту до деякого кластера на певному рівні ієрархічної кластеризації може бути однозначною (кожен даний текст належить лише одному кластеру), або неоднозначною (кожен даний текст може належати декільком кластерам). Кластеризація документів була використана, аби автоматично генерувати ієрархічні кластери документів.

Текстова кластеризація автоматично виявляє групи семантично схожих документів серед заданої великої фіксованої кількості документів.

Слід зазначити, що групи формуються лише на основі попарної подібності описів документів, і жодні характеристики цих груп не задаються заздалегідь, на відміну від текстової класифікації [4].

На кожному робочому комп'ютері існує величезна кількість тек, у яких досить часто зберігається велика кількість документів, які, зазвичай мають абсолютно різну тематику. Людина після певного проміжку часу ледь згадує, що у якій теці знаходиться, а якщо проміжок досягає місяців, то взагалі не може пригадати, у якій теці зберігатися необхідна йому на даний період інформація. Запропонована авторами система дозволяє вирішити цю проблему. Вона автоматично кластеризує документи у теки, які відповідають тематиці документа. Користувачеві необхідно буде скористатися запропонованою системою, і документи віднесуться до логічних за структурою документа тек. У системі на першому етапі документи проходять попередню обробку — скорочення тексту для точнішої класифікації. У нашому випадку препроцессинг (попередня обробка) складається з двох етапів. Документ, що надійшов, попередньо обробляється, перш ніж пройти останні етапи. Спочатку видаляються всі стоп-слова з документу. Стоп- слова — це набір артиклів, таких як: the, a, in, of і так далі. Потім використовується стеммінг — це процес виділення основи слова. Ми вирішили використовувати стеммінг, оскільки він дозволяє максимально скоротити час обробки документа в системі, що, відповідно, веде до оптимізації системи, поліпшення якості її роботи. Ми використовуємо стеммінг алгоритм Портера [5].

Загальна структура цієї моделі даних починається з зображення будь-якого документа як вектора слів, які з'являються в документах набору даних. Вага (зазвичай частоти термів) слів також міститься в кожному векторі. Після попередньої обробки ми використовуємо векторну модель. На сьогоднішній день векторна модель, широко використовується для зображення даних в класифікації і кластеризації документів. Загальна структура цієї моделі даних починається з зображення будь-якого документа як вектор слів, які з'являються в документах набору даних. Вага (зазвичай частоти термів) слів також міститься в кожному векторі. Подібність між двома документами вважається на підставі двох відповідних за властивостями векторів, наприклад Jaccard measure, Euclidean distance та інші. Автори використовували cosine measure. Для опису вакансій, отриманих нашою веб системою, використовуємо метод попереднього оброблення. Попереднє оброблення — це скорочення тексту для точної класифікації. З обробкою методів, різні документи можуть бути створені як структуровані зображення документів. Зазвичай, завдання попереднього оброблення дій включають стандартизацію документа, токенізацию, лематізацию і стеммінг. Технологія цього процесу розглянута на рисунку.



Попереднє оброблення документів

Існує велика кількість різновидів векторної моделі. Ми використовуємо стандартну векторну модель (1)

$$d_{tfidf} = tf_1 \log\left(n/df_1\right),\ tf_2 \log\left(n/df_2\right),\ldots,tf_n \log\left(n/df_n\right). \quad (1)$$

Вважається, що кожен документ є вектором в просторі термів. У його стандартній формі кожен документ зображується вектором частоти (TF) *dtf dtf* = (tf_1,tf_2,.,tf_n), частота *i*-го рядку в документі. Зворотна частота (IDF) документа в частоті документів *i* на log (*n*/*dfi*), де *n* — повне число документів у вибірці, і *dfi* — число документів, які містять *i*-й терм (тобто, частота документа). Нарешті, аби порахувати довжину кожного документа, необхідно кожен вектор документа нормалізувати від 0 до 1 тобто, $\left\| d_{tfidf} \right\|_2 = 1$.

Схожість документів визначається різними методами на підставі двох відповідних за властивостями векторами, наприклад, Jaccard measure і Euclidean distance. У роботі автори використовували cosine measure. Кожен документ зображується як вектор частот слів. Довжиною векторів ми назвемо t. Будуть збережені лише частоти t найбільших слів, що часто зустрічаються. Схожість між двома документами вимірюється як косинус кута між двома векторами з найбільш схожими документами (2). Чим менше кут, тим більше значення косинуса і навпаки, в менш схожих документах, чим більше кут, тим менше значення косинуса. Значення косинуса між двома векторами обчислюється з формули

$$\cos ine\left(d_i, d_j\right) = \frac{\left\langle d_i \bullet d_j \right\rangle}{\left\| d_j \right\|_2 \times \left\| d_i \right\|_2} = \frac{\sum_{t=1}^{t} d_i \times d_j}{\sqrt{\sum_{t=1}^{t} d_i^2} \times \sqrt{\sum_{t=1}^{t} d_i^2}}, \quad (2)$$

де d_i и d_j — вектори документів.

Наступним етапом в нашому підході є ієрархічна кластеризація. Суть ієрархічної кластеризації полягає в послідовному об'єднанні менших кластерів у великі, або розділенні великих кластерів на менші. Ієрархічні алгоритми пов'язані з побудовою дендрограмм (від грецького dendron — «дерево»), які є результатом ієрархічного кластерного аналізу. Дендрограма описує близькість окремих кластерів один до одного, зображає в графічному вигляді послідовність об'єднання (розділення) кластерів.

Автори використовували метод дендрограми для визначення точної кількості кластерів (тек) на основі того, шо кожен об'єкт є окремим кластером. Відстані між цими об'єктами визначаються вибраною мірою [6]. Виникає таке питання — як визначити відстані між кластерами? Існують різні методи об'єднання або зв'язки для двох кластерів. Авторами використано Метод Варда (Ward's method). Як відстань між кластерами береться приріст суми квадратів відстаней об'єктів до центрів кластерів, отримуваний в результаті їх об'єднання (Ward, 1963). На відміну від інших методів кла-

стерного аналізу для оцінки відстаней між кластерами в роботі використано методи дисперсійного аналізу. На кожному кроці алгоритму об'єднуються такі два кластери, які приводять до мінімального збільшення цільової функції, тобто внутрішньогрупової суми квадратів. Цей метод направлений на об'єднання близько розташованих кластерів і «прагне» створювати кластери малого розміру [7].

Наступним етапом є метод *k*-середніх [8]. Загальна ідея алгоритму: задається фіксоване число *k* кластерів і спостереження здійснюються так, що середні в кластері (для всіх змінних) максимально можливо відрізняються один від одного. Даний метод використовується для того, щоб кластеризувати всі неструктуровані (вибрані користувачем) теки з документами по заздалегідь визначеним кластерам (текам), створених на основі дендрограми.

Останнім етапом є повторне використання дендрограми. У середині сформованої теки на основі дендрограми створюється ієрархічна структура вкладених тек.

Висновок

У статті використана векторна модель для класифікації і кластеризації документів. Запропонований підхід застосовано для текстового репозиторія, де використовувалася ітеративна кластеризація, що розділяє на підкластери кожен клас і завершальним кроком була кластерна дендрограмма, для утворення ієрархічної структури цілого набору даних (тек і документів). Комбінований ієрархічний підхід кластеризації документів дозволяє вирішувати проблеми зберігання інформації на комп'ютері і скорочує витрати часу користувача.

СПИСОК ЛІТЕРАТУРИ

1. Ліфшиц Ю. Автоматична класифікація текстів [Електронний ресурс] : [лекція з Data Mining] / Ю. Ліфшиц // Алгоритми для Інтернету : (лекція №6). — Осінь, 2006. — Режим доступу до лк.: http://logic.pdmi.ras.ru/~yura/internet/06ia.pdf — Назва з екрану.

2. Bellot P. Query Length, Number of Classes and Routes through Clusters :Experiments with a Clustering Method for Information. [Електронний ресурс] : (In Proceedings of IEEE ICSC'99) / P. Bellot, M. El-Beze // Springer-Verlag — Berlin, Heidelberg, 1999. — P. 196–205. — Режим доступу до статті: http://wotan.liu.edu/docis/dbl/icscic/1999__196_QLNOCA.htm

3. Zoubin Ghahramani. Unsupervised Learning [Електронний ресурс] : [Data Mining vs Machine learning]: (Machine Learning, Proceedings of the Twenty-Fourth International Conference) / Zoubin Ghahramani // ICML — Corvalis, Oregon, USA — 2007.— Режим доступу до статті:http://www.gatsby.ucl.ac.uk/~zoubin/course05/ul.pdf

4. Lewisand D. Acomparison of two learning algorithms for text categorization [Електронний ресурс] : (In Third Annual Symposium on Document Analysis and Information Retrieval)/ David D. Lewisand, M. Ringuette // 1994. — P. 81—93. — Режим доступу до статті : http://www.research.att.com/~lewis/papers/lewis94b.ps.

5. Porter M. F. An algorithm for suffix stripping [Електронний ресурс] : [Text retrieval] / M. F. Porter // Program – 1980. — №4(3). — P. 130—137. — Режим доступу до статті: http://tartarus.org/~martin/PorterStemmer/def.txt . — Назва з екрану.

6. Everitt B. Cluster Analysis [english] / B. Everitt. — NewYork : Wiley,1993. — 283 p. — Heinemann Educational Books LTD. — Бібліогр. в підрядк. Прим. — ISBN 034057237X / 9780340572375 / 0-340-57237-X

7. Чубукова І. А. Методи кластерного аналіза. Ієрархічні методи [Електронний ресурс]: (INTUIT.ru::Інтернет-Університет Інформаційних Технологій. Дистанційна освіта. — 2003-2008)/ І. А. Чубукова // Data Mining : (лекція № 13). — 2006. — Режим доступу до лк.:http://www.intuit.ru/department/database/datamining/13/2.html

8. Bradley, P. S. Constrained k-means clustering [Електронний ресурс] / Bradley, P. S., Bennett, K. P. Demiriz, A. // Microsoft Research. MSR-TR-2000-65. 2000. — Redmond, W. A. — Режим доступу до статті.: http://www.litech.org/~wkiri/Papers/wkiri.html

Рекомендована кафедрою інтелектуальних систем

Надійшла до редакції 8.09.08
Рекомендована до друку 20.10.08

Шатовська Тетяна Борисівна — доцент кафедри програмного забезпечення електронних обчислювальних машин, ***Каменєва Ірина Віталіївна*** — студентка.

Харківський національний університет радіоелектроніки