УДК 004.93(015.7)

АМОНС О.А.,
ЯНОВ Ю.О.,
БЕЗПАЛИЙ І.О.

КЛАСТЕРИЗАЦІЯ ДОКУМЕНТІВ НА ОСНОВІ СТАТИСТИЧНОЇ БЛИЗЬКОСТІ ТЕРМІВ

У статті описано підхід до кластеризації колекції документів з невідомою наперед кількістю кластерів. В основу підходу покладено метод, оснований на статистиці появи ключових термів. Запропоновано модифікацію методу знаходження матриці подібності на основі схожості косинуса. Для аналізу якості й знаходження граничних значень алгоритму використана модифікація функції конкурентної подібності. Підхід реалізований у вигляді прикладного застосування сервера SmartBase. Наведені результати експериментальних досліджень запропонованого підходу до кластеризації інформації з використанням часто вживаного текстового корпусу підтверджують працездатність запропонованих рішень.

In the given work the approach to clustering of documents collections with unknown quantity of clusters is described. A method of finding matrix of similarity is improved. The method is based on the statistics of key terms occurrence in documents. For quality analysis and finding of limiting values of algorithm, there was used a function of competitive similarity improving. The approach is realized as the application server SmartBase's application. Implementation details and results of the process are shown. Russian text set is used.

Вступ

Зі швидким ростом всесвітньої павутини, впровадженням у міністерствах, відомствах і великих організаціях різноманітних засобів підтримки інформаційно-аналітичної діяльності, насамперед інформаційних систем, систем електронного документообігу, інформаційні бази інформаційно-телекомунікаційних систем спеціального призначення (ІТС СП) набувають суттєвих розмірів. Вони стрімко розширюються переважно за рахунок неструктурованих даних, що породжує проблему швидкого орієнтування в них. Дійсно, відсутність можливості отримувати актуальну і повну інформацію з конкретної теми, яка цікавить користувача, перетворює в непотріб велику частину накопичених ресурсів і робить марними зусилля фахівців. Оскільки безпосередній пошук і анналіз інформації за заданою темою досліджень вимагає все більших трудовитрат, багато рішень приймаються на основі неповного бачення проблеми. Використання засобів для автоматичної систематизації і класифікації текстової інформації дозволяє скоротити час на пошук потрібної інформації, забезпечити її повноту і тим самим підвищити якість дослідження та швидкість реагування на зовнішні зміни. Одним із засобів для підвищення ефективності роботи фахівців в умовах функціонування ІТС СП є інструменти кластеризації текстів, які підлягають аналізу. Стаття присвячена аналізу існуючих методів і засобів пошуку текстової інформації, розробленню дієвого алгоритму кластеризації текстової інформації і його дослідженню на наявному текстовому корпусі.

1. Пошук документів і кластеризація

Пошук документів є однією з найуживаніших операцій оброблення інформації. Дійсно, перш ніж документи можуть бути опрацьованими різними процедурами, їх необхідно знайти серед величезного числа подібних. Якщо пошук документів буде не досить ефективним, то й загальні процедури оброблення інформації, представленої у вигляді документів, не будуть ефективними. Якщо історію пошуку документів можна розпочинати від зародження бібліотечних інформаційних систем [1], то період його інтенсивних досліджень започаткований формуванням Глобальної мережі і її перетворенням у практичний інструмент підтримки різних видів діяльності людини. Саме тоді з'явилися відомі пошуковики, ефективність яких швидко зростала під впливом нагальних потреб [2]. Останні події зі світу пошуку документів свідчать, що для збільшення його ефективності застосовуються не лише моделі, традиційні для морфологічного і синтаксичного рівнів мовної системи, але й семантичні моделі [3].

Хоч сутність проблеми пошуку документів розуміють як користувачі, так і фахівці в галузі

інформаційних технологій (ІТ), лише останні розуміють її складність як науково-практичної проблеми. І джерело цієї складності не стільки у необхідності застосування серйозних моделей і методів лінгвістики, скільки у необхідності такої їх реалізації засобами ІТ, яка задовольняє високий рівень вимог до пошуку, насамперед його адекватності і швидкості. Для кращого розуміння проблем пошуку наведемо необхідні визначення. Під адекватністю розуміють відповідність результатів пошуку інформаційним потребам користувача з погляду визначених цілей і створених на їх основі запитів. Швидкість розглядається у контексті ефективності алгоритму і можливого часу його виконання. Сам пошук інформації становить собою процес виявлення в деякій множині (колекції) документів (текстів) усіх таких з них, які присвячені певній темі, задовольняють заздалегідь визначеній умові пошуку (запиту) чи містять необхідні (що відповідають інформаційній потребі) факти, відомості, дані. Кластеризація – це процес створення кластерів. Згідно з розповсюдженним визначенням, "кластери – це неперервні області (якогось) простору з більш високою щільністю елементів, відділені від інших таких ж областей з відносно низькою щільністю елементів". Під кластеризацією документів розуміють процес поділу всієї колекції на групи, у середині яких знаходяться близькі за тематикою документи, а в різних групах, навпаки, далекі.

Хоч існують різноманітні підходи до побудови пошуковиків, для найефективніших з них загалом властиве застосування, крім зазначених моделей і методів лінгвістики, технологій кластеризації та індексування. І якщо індексування швидше пов’язане з технологічними аспектами, то кластеризація – це традиційна для ІТ наукова проблема. На сьогодні, як засіб для впорядкування текстової інформації, у пошукових системах застосовують різні методи кластерного аналізу. Це, насамперед, статистичні класифікатори на основі ймовірносних методів. Найбільш відомими з них є сім’я Байесових алгоритмів. Наступною групою є класифікатори, що використовують методи на основі штучних нейронних мереж, наприклад алгоритми ART, SOM [4]. І останню групу складають класифікатори, що базуються на функціях схожості, насамперед k-means, Forel, FRiS Cluster [5,6,11].

Одними із найуживаніших є методи, засновані на метриці близькості. У цьому випадку документи представляються у вигляді вектору ознак у просторі ознак, тобто набором ключових слів. Є декілька підходів до його формування. В найпростішому випадку кожна ознака відповідає присутності в тексті одної із словоформ, що зустрічається у текстовій колекції. Величину кожного елемента вектора можуть підраховувати по різному: наприклад, прирівнювати одиниці, якщо ознака зустрічається у даному тексті, чи нулю у іншому разі; вона може бути рівною кількості входжень його у документ, нормованою до кількості ознак; чи також враховувати частоту появи ознаки у всьому текстовій колекції. Таке представлення має суттєвий недолік – простір ознак має велику розмірність, більша частина його елементів є надмірними, навіть шкідливими. Для подолання цієї проблеми використовуються методи зменшення розмірності простору ознак. Це, насамперед, виділення лем слова, нормальних форм, видалення стоп-слів, використання синонімічних груп тощо. Але в цьому випадку є ймовірність не використати значущу інформацію.

На наступному кроці підраховується матриця близькості між векторами документів. І виконується власне кластеризація. Більш детальну інформацію з описом особливостей існуючих алгоритмів можна знайти в працях [7,8].

Крім загальних проблем, для всіх методів кластеризації текстів постає проблема відображення змісту кластеру, на основі якого до нього приєднується той чи інший текстовий документ. Це необхідно для зручного використання результатів людиною. Найбільш розповсюджений підхід до рішення цієї проблеми складається у використанні представлення кластеру у вигляді набору найбільш важливих слів.

2. Постановка проблеми

Проблема полягає у розробленні текстового кластеризатора і дослідженні його властивостей. Загальну модель текстового кластеризатора можна представити багатоосновною алгебраїчною системою такого вигляду:

$$R = < T, C, D, B, R, S > \qquad (1)$$

де:

$T = \{T_1, T_2, ..., T_1, ..., T_n\}$ – множина текстів, що підлягають класифікації (колекція);

$T_l = \{t_1, t_2, ..., t_j, ...\}$ – множина термів, з яких складається l-й документ;

$C = \{C_1, C_2, ..., C_k\}$ – множина класів-рубрик (кластерів), де k – кількість кластерів;

$D = \{D_1, D_2, ..., D_m\}$ – множина описів, кожний з яких має певну внутрішню структуру, де m – кількість описів;

$B = \{b_1, b_2, ..., b_k\}$ – множина еталонних зразків (стовпів), $i = 1,2,\ldots, k$;

$R \subset C \times D$ – відношення між кластерами і описами, яке має таку властивість: $\forall C_i \in C \; \exists \; D_j \in D : (C_i, D_j) \in R$, причому кожному кластеру відповідає єдиний опис;

S – сигнатура, яка включає такі операції:

S_1: $T \rightarrow C$ – операція кластеризації, яка полягає у виконанні перетворень над текстами, після яких, або робиться висновок про належність документа T_l зі структурою D_l до класу C_i, або висновок про створення нового кластеру C_j, до якого можна буде віднести даний текстовий документ. Будемо вимагати, щоб жодний текст не міг відноситись до декількох кластерів одночасно;

S_2: $C \times C \rightarrow C$ – теоретико-множинна операція перетину кластерів;

S_3: $C \times C \rightarrow C$ – теоретико-множинна операція об'єднання кластерів.

3. Загальний опис підходу до розв'язання проблеми

Підхід, реалізований авторами, полягає у використанні векторної моделі текстових документів, згідно з якою кожний текстовий документ представляється у вигляді вектору зважених ознак і належність документів до кластерів визначається на основі міри близькості відповідних векторів.

Досить логічно кластеризація виконується у таких чотири основних етапи:

Етап 1. Це підготовчий етап, який полягає у переході від множини текстів T до множини T^* їх векторів.

Етап 2. На цьому етапі відбувається перехід до множини описів документів з урахуванням ваги ознак. Будемо використовувати статистичні міри ваги, які добре зарекомендували себе у пошуку документів, насамперед міри, що характеризують частку деякого терму документа у загальній кількості термів та їх появи в документах всього набору [8].

Етап 3. Будуємо матрицю близькості між документами на основі популярної функції схожості косинуса, фізичним змістом якої є косинус кута між векторами.

Етап 4. На цьому етапі виконується власне кластеризація.

Розглянемо наведені етапи більш детально.

1) Підготовчий етап. Для кожного документу T_l множини текстів T виділяємо множину значимих слів, які приводяться до нормальних форм. Будемо традиційно називати ці слова ключовими термами і позначати t_q. Як і в багатьох інших підходах, ми розглядаємо в якості ключових ті терми, частота яких у даному тексті істотно перевищує деяку середню частоту. Крім того ми будемо виключати стоп-слова, такі як прийменники, сполучники тощо.

2) Від колекції текстів переходимо до множини описів текстів. Знаходимо вагу кожного терму. Для цього кожному терму t_i документу T_l ставимо у відповідність статистичну міру w_{li}, що характеризує відношення кількості входжень цього терму у документ до загальної кількості термів та враховує частоту появи терму в документах всієї колекції:

$$w_{li} = -\log(p(t_i) \cdot f_t(T_l, t_i)) \qquad (2)$$

$$f_t(T_l, t_i) = \frac{f_r(T_l, t_i)}{f_r(T_l, t_i) + 1 + \frac{d(T_l)}{350}},$$

$$p(t_i) = 1 - e^{-1.5 \cdot \frac{f_c(t_i)}{n}}$$

де

$f_r(T_l, t_i)$ – число входжень терма t_i у документ T_l;

$d(T_l)$ – кількість термів у документі T_l;

$f_c(t_i)$ – число входжень терма t_i в колекцію.

3) Будуємо матрицю близькості між документами на основі популярної функції схожості косинуса, фізичним змістом якої є косинус кута між векторами:

$$sim(T_1, T_2) = \frac{\sum_i w_{1i} w_{2i}}{\sqrt{\sum_i w_{1i}^2 \sum_i w_{2i}^2}}, \qquad (3)$$

де w_{1i}, w_{2i} - вага терма ti в документах T1,T2 відповідно.

У виразі (3) враховуються лише терми, що входять одночасно як в документ T_1, так і в до-

кумент T_2. Але цій оцінці притаманні певні недоліки. Продемонструємо їх на прикладі. Розглянемо два тексти:

a) "Президент обратился с представлением в Конституционный суд с просьбой растолковать конституционный механизм замены отдельных членов правительства."
b) "Как Президент выражаю почет и благодарность всем ветеранам трагической афганской войны и украинским воинам-миротворцам за отвагу, доблесть, верность присяге и чести"

У цих текстах є лише одне спільне ключове слово "Президент" і воно вживається лише одного разу. І оскільки кількість ключових термів у кожному з документів приблизно рівна, то використання виразів (2) і (3) дає оцінку близькості $\mathrm{sim}(T_1,T_2) \sim 1$, що не повною мірою характеризує дійсний зв'язок цих текстів.

Цього недоліку можна позбутися використовуючи для оцінки близькості такий вираз:

$$sim(T_1,T_2) = \frac{\sum_i w_{1i} w_{2i}}{\sqrt{\sum_i w_{1i}^2 \sum_i w_{2i}^2}} \cdot \sum_i \frac{w1_{1i} + w2_{2i}}{w1_{1i} + w2_{2i}}, \qquad (4)$$

$w1_{1i}, w2_{1i}$ - вага термів, що одночасно присутні в обох документах.

4) Власне кластеризація. Будемо виконувати кластеризацію на основі FRiS-функцій [10]. Розглядаємо колекцію T із n документів, що розбита на k угрупувань. Кожне угрупування i описане одним еталонним об'єктом (стовпом, центроїдом) b_i. Для будь-якого документа $T_1 \in T$ можна знайти відстань $r(T_1,b_i)$ до найближчого стовпа угрупування i. Тоді $r1(T_1,b_i) = \min_i r(T_1,b_i)$ - близькість до найближчого стовпа i, а $r2(T_1,b_i) = \min_{i \neq i^*} r(T,b_i)$ - близькість до найближчого конкурента.

FRiS-функція [10], модифікована для використання матриці близькості між документами, має такий вигляд:

$$F(T,B) = \left(\frac{r1(T,b) - r2(T,b)}{r1(T,b) + r2(T,b)} \right) \qquad (5)$$

Вона є мірою подібності об'єкту T зі стовпом b_i у конкуренції з іншими стовпами.

Знайшовши середнє значення FRiS-функції по усій вибірці, за допомогою виразу (6) отримаємо величину $F(B)$, що характеризує наскільки повно набір стовпів характеризує колекцію:

$$F(B) = (1/m) \sum_{Ti \in T} F(T_i,b) \qquad (6)$$

У роботі також будемо використовувати редуковану FRiS-функцію, у якій стовп конкурента рівновіддалений від кожного об'єкта на відстань $r2^*$. Тоді вирази (5), (6) набудуть такого вигляду:

$$F^*(T,B) = \left(\frac{r1(T,b) - r2^*(T,b)}{r1(T,b) + r2^*(T,b)} \right) \qquad (7)$$

$$F(B) = (1/m) \sum_{Ti \in T} F^*(T_i,b) \qquad (8)$$

4. Опис алгоритму кластеризації

Деталізуючи наведений узагальнений опис, кластеризацію будемо виконувати за допомогою наступного алгоритму:

Крок 1. При k=1 створюємо новий кластер C_{first} і всі документи включаємо до нього.

Крок 2. У кластері, отриманому на попередньому етапі, призначаємо стовпом довільно вибраний документ t і для нього по формулі (8) обчислюємо середню редуковану FRiS-функцію $F^*(B)$.

Крок 3. Крок 1 повторюється при призначенні стовпами всіх m об'єктів кластеру почергово. Першим стовпом b_1 обирається документ T^*_1, для якого величина F^* виявляється максимальною.

Крок 4. Призначаємо другим стовпом довільно обраний документ і підраховуємо для нього середню редуковану FRiS-функцію.

Крок 5. Крок 4 повторюється при призначенні стовпами всіх m документів кластера почергово, крім існуючого стовпа b_1. Другим стовпом b_2 вибирається документ T^*_2, для якого величина F^* виявляється максимальною.

Крок 6. Замість одного кластеру C_{first} створюємо два нових з центроїдами b_1 та b_2.

Крок 7. Після того, як були знайдені два нових стовпи, уся колекція розподіляється за наступним правилом: документ відноситься до того кластеру, для якого близькість $r1$ до найближчого центроїду максимальна.

Крок 8. Якщо при k=1 у перший кластер C_{first} входили всі m документів, то тепер при k=2 документи розподіляються між двома новими кластерами. При цьому може вийти так, що для опису кластера C_1 найкращим виявиться не стовп b_1, а якийсь інший документ із цього кластера. Для покращення місця розташування

стовпа b_1 виконаємо наступну процедуру. Почергово для кожного кластера, у нашому випадку для C_1,C_2, призначаємо внутрішні документи на роль стовпа, і за допомогою виразу (6) обчислюємо середню FRiS-функцію $F(T_{1i},b_i)$. Центроїдом вибирається той документ, що забезпечує максимальну величину FRiS-функції. Аналогічно у кластері C_2 визначається нове положення стовпа b_2 на основі максимуму функції $F(T_{2i},b_i)$. На цьому етапі замість редукованої FRiS-функції використовуємо звичайну, що відображає процес конкуренції між реальними стовпами.

Крок 9. Для подальшого розбиття сукупності документів, необхідно вибрати кластер з мінімальним загальним відхиленням від середнього значення сумісної близькості документів. Підрахунок її виконуємо за допомогою виразу (9)

$$E_k = \sqrt{\sum_{i=1}^{nk} (sim(T_i, b_k) - middleSim_k)^2}, \quad middleSim_k = \frac{1}{n_k} \sum_{i=1}^{nk} sim(T_i, b_k) \quad , \quad (9)$$

де n_k – кількість документів у кластері.

Кластер з мінімальною сумісною близькістю E_k вибираємо для подальшої роботи.

Крок 10. Процес продовжується, переходячи на крок 2, тобто у вибраному кластері шукаємо 2 стовпа за допомогою описаного вище алгоритму, створюємо 2 нових кластера, перерозподіляємо всі документи між всіма стовпами, і уточнюємо місце розташування центроїда. Процес повторюється доти, поки колекція документів не поділиться на k кластерів.

5. Програмна реалізація

Реалізація і дослідження розробленого алгоритму виконується у рамках розробки системи прийняття рішень на платформі SmartBase.

Для виконання досліджень була використана колекція документів з Інтернет-ресурсу аналітичної газети «Дзеркало тижня» за 3 періоди. Вона включає новини з 26 різних рубрик. Рубрики використовувались як еталони у подальшому дослідженні.

Для стемінгу, тобто виділення нормальної форми слова та виявлення належності його до тієї або іншої частини мови, було використано вдосконалений алгоритм SnowBall.

У виконаних експериментах використовувався список стоп-слів. Він зчитувався з файлу stopword.xml, створеного на основі списку Штейнфельдта, а також з переліку слів, вільно розповсюджуваного компанією "Яндекс" продукту Yandex.Server-FREE-020-3.8.3. Список включає 279 часто вживаних слів російської мови.

Для видалення HTML-тегів була розроблена власна бібліотека HTML Parser, що надає потрібну функціональність.

Програмний інтерфейс складається з web-застосування, яке надає можливість зчитувати новини з сайту http://www.zn.ua/ за певний період чи з локального файлу даних. Вхідними даними для алгоритму є максимальна кількість кластерів k, $k \in (1,...,n)$, де n – кількість документів у колекції. У алгоритмі кластеризації використовується лише тіло документу (основний текст), заголовок і додаткова інформація не враховуються.

Експерименти проводились на двопроцесорному сервері Intel(R) Xeon (TM) CPU 2,40 GHz 2,40 GHz 1GB RAM. Результати передавались клієнту по http протоколу. Характеритики клієнта наступні: Intel(R) Core(TM) 2 Duo CPU 2,33 GHz 2GB RAM.

6. Методика проведення експерименту

Кожний документ із колекції методом експертної оцінки було віднесено до певних рубрик (наприклад, людина, внутрішня політика, міжнародна політика, право, тощо). Максимальна кількість таких рубрик 26. Але для кожної конкретної колекції документів рубрик може бути менше. На основі використання цих даних були створені еталонні кластери.

Існує декілька характеристик оцінки якості роботи текстового класифікатора. Найбільше поширення одержали точність (P) і повнота (R). Вони також використовуються при оцінці якості пошуку за запитом, наприклад, у пошукових машинах мережі Інтернет.

Під правильною і неправильною рубрикацією будемо розуміти випадки, коли класифікатор приписує аналізований документ деякій рубриці, що розцінюється деяким експертом як вірне і невірне рішення відповідно. Під невірнім відсіюванням документу розуміємо випадок, коли класифікатор не приписує документ рубриці, що, на думку експерта, невірно.

Для аналізу отриманих результатів використовувалися графіки точності/повноти. Незважаючи на широку поширеність і популярність метрик точності і повноти у задачах інформа-

ційного пошуку, стосовно задачі розбиття документів на кластери їх застосування вимагає деяких уточнень. Введемо наступні позначення:

C_{ei} – i-й нетривіальний (утримуючий більше одного документа) еталонний (складеними експертами) кластер;

C_j – j-й нетривіальний текстовий (побудованою системою) кластер;

$$M_{ij} = \begin{cases} \left|C_{ei} \cap C_j\right|, при \left|C_{ei} \cap C_j\right| > 1; \\ 0, у \quad іншому \quad випадку; \end{cases}$$

$$I_{ij} = \begin{cases} 1, при \left|C_{ei} \cap C_j\right| > 1; \\ 0, у \quad іншому \quad випадку; \end{cases}$$

$$U_j = \begin{cases} \sum_i I_{ij}, при \sum_i I_{ij} > 1; \\ 0, у \quad іншому \quad випадку; \end{cases}$$

$$K_i = \begin{cases} \sum_j I_{ij}, при \sum_j I_{ij} > 1; \\ 1, у \quad іншому \quad випадку. \end{cases}$$

Тоді точність є відношенням різниці потужності перетину документів в еталонних і в тестових кластерах і кількості повторно використаних (при побудові даного перетину) еталонних кластерів до ефективної кількості документів у тестових кластерах:

$$P = \frac{\sum_{ij} M_{ij} - \sum_j U_j}{\sum_i \left|C_{ei}\right| \cdot K_i}$$

Повнота виражається відношенням різниці потужності перетину документів в еталонних і в тестових кластерах і кількості повторно використаних (при побудові даного перетину) еталонних кластерів до кількості документів в еталонних кластерах:

$$R = \frac{\sum_{ij} M_{ij} - \sum_j U_j}{\sum_j \left|C_j\right|}$$

7. Результати експериментальних досліджень

Дослідження проводились на трьох колекціях документів, за різний проміжок часу:

- за період з 29.12.2007 по 26.04.2008, 1515 документів;
- за період з 26.04.2008 по 27.00.2008, 1543 документів;
- за період з 31.05.2008 по 5.06.2008, 418 документів;

У тестуванні брало участь 2 алгоритми: розроблений авторами алгоритм Cluster і алгоритм FRiS Cluster [11] (далі FRiSCluster). Останній обраний за результатами досліджень, наведених у праці [9], як кращий з існуючих алгоритмів цього класу. Обидва алгоритми оперують поняттям центра кластеру. Після закінчення своєї роботи вони надають користувачу не тільки результуючий розподіл, але і еталонні зразки - стовпи.

На рисунках 1 - 3 наведені результати оцінки якості кластеризації за показником точності/повноти для різної кількості документів у колекції новин, від 100 до 1600. Кожний документ було представлено за допомогою двох десятків найбільш інформативних термів.

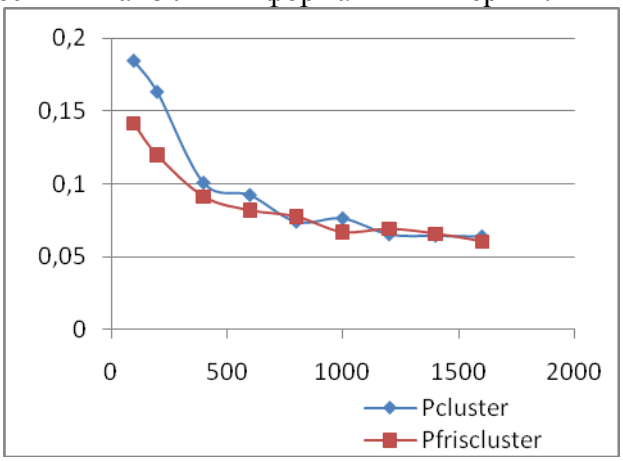

Рис. 1.Порівняння точності алгоритмів

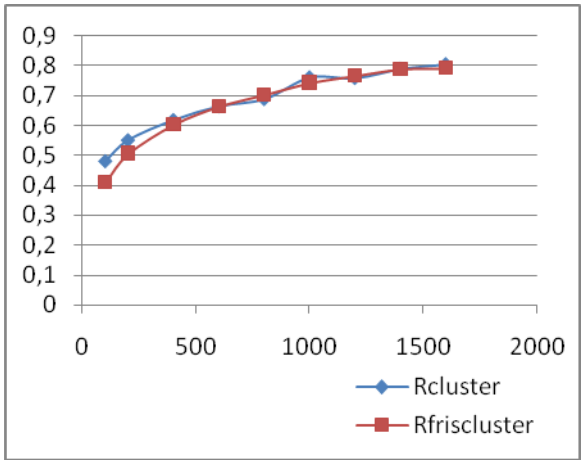

Рис. 2. Порівняння повноти кластеризації

Графіки точності і повноти, зображені на рисунках 1 і 2 відповідно, показують що розроблений алгоритм має співставні з конкурентом, який є одним із найкращих серед відомих алгоритмів, результати. Тільки при невеликій кіль-

кості документів його точність відчутно перевищує точність алгоритму FRiSCluster

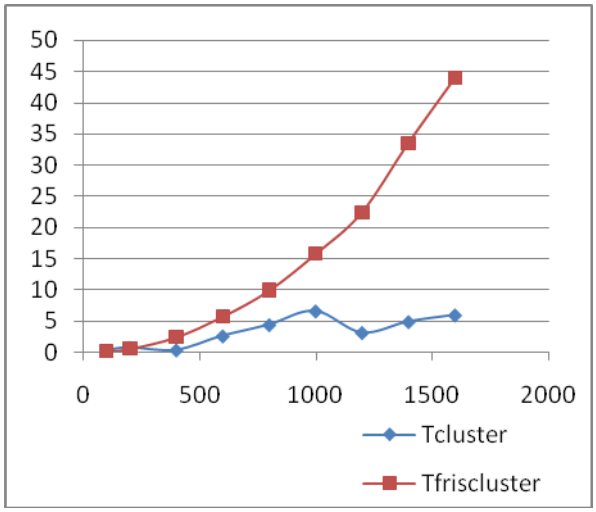

Рис. 3. Порівняння швидкості роботи алгоритмів

На рисунку 3 відображено залежність часу роботи алгоритмів у секундах (ордината) від кількості документів (абсциса). Судячи з графіків, алгоритм Cluster має суттєві переваги, час його роботи менш залежить від збільшення кількості документів у колекції.

Необхідно звернути увагу на невелику точність результатів. Це пояснюється насамперед недосконалістю алгоритму стемінгу, не врахуванням віднесення слова до тієї або іншої частини мови, не врахуванням синонімії, тобто браком засобів виявлення значущих слів та їх відношень.

8. Перспективи досліджень

Для поліпшення показників якості роботи прикладного застосування і розширення сфери його використання видається необхідним виконати такі роботи:

1. Розробка, реалізація і дослідження більш ефективного алгоритму стемінгу.
2. Розробка, реалізація і дослідження моделей представлення семантики і методу кластеризації колекції документів на основі статистичної і семантичної близькості, що дозволить врахувати синонімію та інші відношення термів.
3. Розробка, реалізація і дослідження алгоритму виявлення словосполучень.
4. Розширення кількість мов, з якими працює текстовий кластеризатор.

Висновки

5. Запропоновано підхід до кластеризації колекції документів з невідомою наперед кількістю кластерів на основі статистичних показників, що характеризують кількість входжень ключових термінів у документах і колекції у цілому.
6. Удосконалено метод знаходження матриці подібності на основі схожості косинуса, для аналізу якості й складності якого використана модифікація функції конкурентної подібності.
7. Підхід реалізований у вигляді прикладного застосування сервера SmartBase і виконане його експериментальне дослідження з використанням наявного текстового корпусу. Результати досліджень підтверджують працездатність запропонованих рішень і їх відповідність за показниками точності і повноти відомим розробкам, а за показником швидкості перевищують їх.

Перелік посилань

1. Дж Солтон. Динамические библиотечно-информационные системы. М.: - Мир, 1979.- с.557.
2. Козлов Д.Д. Проблемы применения методов поиска тематических сообществ к задаче тематического информационного поиска в интернет // Труды Всероссийской научной конференции "Методы и средства обработки информации" -М.: Издательский отдел факультета ВМиК МГУ, 2003. - С. 211-215.
3. AllaZaboleeva, Yulia Orlova Computer-aided system of semantic text analysis of a technical specification//Information Technologies and Knowledge. – 2008. – Vol.2. – P.139-145.
4. Vassilis G. Kaburlasos Unified Analysis and Design of ART/SOM Neural Networks. Heidelberg: - Springer Berlin, Volume 4507, 2007.-p 80-93
5. MacQueen J. Some methods for classification and analysis of multivariate observations // Proceedings of the 5th Berkley Symposium on Mathematical Statistic and Probability, University of California Press, 1967, Vol.1, p. 281-297.
6. Peter Grabusts A Study of Clustering Algorithm Application In RBF Neural Networks. //Information technology and management science. - Riga, - 2001. - 5.serija., p.50-57.
7. Корунова Н. В., Кластеризация документов проектного репозитария на основе нейронной сети Кохонена // Труды конф. «Нечеткие системы и мягкие вычисления». – М. - 2008. – С. 77-86.

8. Киселев М. В., Пивоваров В. С., Шмулевич М. М. Метод кластеризации текстов, учитывающий совместную встречаемость ключевых терминов, и его применение к анализу тематической структуры новостного потока, а также ее динамики// Автоматическая обработка веб-данных. - М., 2005. - С. 412-435.
9. Красильников П.В. Воспроизведение лучших результатов ad hoc поиска семинара РОМИП. - М.: IMAT, 2007. – 220с.
10. Борисова И.А., Дюбанов В.В., Загоруйко Н.Г., Кутненко О.А. Использование FRiS-функции для построения решающего правила и выбора признаков // Материалы Всероссийской конференции с международным участие «Знания – Онтологии – Теории» (ЗОНТ–07), Новосибирск, 2007 г. – Т. 2. – С. 67-76.
11. Борисова И.А., Загоруйко Н.Г., Функции конкурентного сходства в задаче таксономии //Материалы Всероссийской конференции с международным участие «Знания – Онтологии – Теории» (ЗОНТ–07), Новосибирск, 2007 г. – Т. 2. – С. 77-86.