УДК 004.912

С. В. Ковальчук, аспірант
Одеський національний політехнічний університет, Україна
serhiy_kovalchuk@mail.ua

Вдосконалення методу індексації термінів в згрупованих документах для реалізації пошуку в корпоративних системах

Розроблено метод індексації термінів, який поєднаний з процесом їхнього виділення. Запропоновано формувати індекс з чотирьох компонентів: ідентифікатора документу, номера речення в документі, позиції початку терміну в реченні та кількості слів в терміні. Індексація дозволяє виконати відповідно до запиту користувача не тільки пошук документу, але і його фрагменту.

Ключові слова: термін, корпоративна система, індексація, інвертований індекс, словник предметної області.
DOI: 10.31474/1996-1588-2018-2-27-17-22

Постановка проблеми

Знаходження актуальної для користувача інформації є головною метою будь якої інформаційної системи. Словник предметної області дає можливість зберігати терміни усіх документів корпоративної системи, цим самим спрощує процес роботи над пошуком необхідної інформації. Одне з можливих призначень словника предметної області – це оцінка конкретного документу з точки зору розміщеної в ній інформації. Оскільки процес виділення термінів пов'язаний з послідовним аналізом всіх речень та слів базового тексту, з'являється інтерес поєднати цей процес з визначенням місця розташування терміну, тобто ввести їхню індексацію. Сучасні потужності комп'ютерної техніки дозволяють виконувати прямий прохід по документах з цілю пошуку потрібної інформації. Однак індексація дозволить більш якісно і швидше знаходити матеріал з орієнтацією не тільки на весь документ, але і на його фрагменти. Подібні рішення особливо актуальні для корпоративних систем, де при відносно невеликих обсягах інформації, що зберігається, потрібно швидкий і якісний пошук.

Аналіз останніх досліджень і публікацій

Спроби здійснити індексацію термінів під час їхнього виділення в документах до словника предметної області було описано в роботі [1]. Процес виділення фрагментів тексту було висвітлено в роботі [2]. Але не було описано як правильно формувати та розташовувати індекс в словнику предметної області. Проблема швидкого і якісного пошуку різних термінів в документах пов'язана з можливістю зберігання інформації про розташування терміну з урахуванням кількості наявних в ньому слів. Існує безліч підходів до індексації в пошукових системах [3]. Кожен з них має певні переваги і недоліки. Вибір способу безпосередньо залежить від поставлених завдань. Одним з найпоширеніших є інвертований (зворотний) індекс [4]. Його реалізація передбачає варіанти з формуванням списків, в кожному з яких знаходиться термін і список документів, в яких він зустрічається. Даний індекс можна розширити, додавши до нього позиції слів в відповідних документах. При розширені, може виникнути проблема великого розміру сформованого інвертованого індексу, що не дасть змогу повністю розмістити його в оперативну пам'ять, для швидкого пошуку. Ця проблема не поширюється на корпоративні мережі з відносно невеликою кількість документів в порівнянні з мережею Інтернет. Процес пошуку за допомогою індексів було розглянуто в роботі [5], а особливості використання індексації та необхідний об'єм пам'яті для зберігання та використання індексів висвітлено в роботі [6]. Не зважаючи на це, залишається не вирішена проблема щодо об'єднання процедури виділення термінів та подальшої їхньої індексації.

Постановка задачі

Для здійснення процесу індексації термінів та кластеризації документів, потрібно вирішити наступні проблеми:

1. Сформувати позиції термінів в документі.
2. Визначення частоти появи терміну в документі.
3. Визначення об'єму індексу та представлення його в словнику предметної області.
4. Реалізувати метод кластеризації документів.

В результаті вирішення виділених проблем буде реалізовано метод індексації термінів, який дасть змогу спростити процес подальшого пошуку інформації в корпоративних системах. Також буде реалізовано метод кластеризації документів, для прискорення процесу пошуку.

Мета дослідження

Головною метою дослідження є скорочення часу на пошук необхідної інформації за рахунок індексації термінів в корпоративних системах.

Індексація термінів для пошуку інформації

Характеристики документа по множині вхідних до нього термінів і частоти їх появи досить для визначення загальної тематики документа. Але часто користувачеві потрібно знайти фрагмент документа, що відповідає більш вузькій темі. В

цьому випадку зазвичай використовують індексацію.

Побудова системи інформаційного пошуку, яка базується на індексації, залежить від обсягу інформації, яку потрібно обробляти і від потужності комп'ютерного забезпечення. Способи побудови індексу можна розділити на дві категорії. Перша категорія використовує оперативну пам'ять (весь індекс розміщується в оперативній пам'яті), інша - дискову пам'ять [5].

Словники предметних областей має сенс створювати в рамках деякої організаційної системи. Можливо, це буде корпоративна система. У цих умовах кількість документів зазвичай не перевищує 100000, а середня довжина документу становить 300 речень [6]. В таких умовах для визначення місця розташування терміну пропонується використовувати інвертований індекс, що розміщується в оперативній пам'яті. Це забезпечить простоту реалізації індексації та швидкий пошук терміну.

Обсяг пам'яті, яку використовує індекс текстового пошуку, в значній мірі визначається складністю тексту документа. При цьому між розміром індексу текстового пошуку та розміром вихідних даних існує приблизно лінійна залежність. Як правило, розмір індексу на диску становить від 50% до 150% від розміру початкового тексту [6]. Крім вільного простору для оновлення індексу і великої кількості службової інформації, індекс може становити в середньому 25% від загального розміру документів. Відношення розміру індексу текстового пошуку до розмірів вихідних документів залежить від середнього розміру документів, кількості і розподілу унікальних термінів.

Середню довжину речень можна визначити за допомогою індексу Флеша *FRE* [7]. Тексти корпоративних систем можна вважати середньої складності для сприйняття. У цьому випадку індекс Флеша становить $FRE = 65$. Це означає, що середня довжина речень становитиме 20 слів. Середній розмір документів корпоративних системах становить 300 речень, а кількість документів - може досягати 100 тисяч [6].

Загальний обсяг текстової інформації в словах матиме вигляд:

$$V = ASL * SCount * TCount, \quad (1)$$

де ASL – середній розмір речень в словах; $SCount$ – середня кількість речень в документі; $TCount$ – середня кількість документів в корпоративній системі.

Розмір текстової інформації документів, у розмірі 300 речень і 20 слів кожне, може становити 40 Кб. Таким чином, розмір усіх текстових документів корпоративної системи в середньому може досягати 4 Гб. [6].

Для кожного документу створюється тимчасовий документ (метадокумент), в якому кожне речення записано з нового рядка. Тимчасові документи представлені в форматі txt. У порівнянні з документами формату docx, txt займають вдвічі менше пам'яті.

Індексування відбувається під час знаходження усіх можливих варіантів представлення терміну tm, де tm – множина варіантів представлення терміну. Для кожного варіанту формується відповідна позиція. Множина позицій для конкретного tm має вигляд:

$$tmPos = \{tPos_p\} p = 1, count, \quad (2)$$

де $tPos_p$ – позиція терміну в документі; $count$ – кількість представлень терміну у множині tm.

Кожна позиція $tPos_p$ знайденого терміну характеризується атрибутами:

$$tPos_p = < DocId, Nr, pos, len >, \quad (3)$$

де $DocId$ – індекс документу, в якому знаходиться термін; Nr – номер речення, в якому знаходиться термін; pos – номер символу, з якого розпочинається термін (позиція терміну в реченні); len – кількість символів в терміні (довжина терміну).

Якщо множина tm має більше одного представлення терміну t, тоді інформація про індекс документу $DocId$ та номер речення Nr визначається тільки для першого представлення терміну $t_1 \in tm$, оскільки вона є спільною для усіх термінів, виділених в даному реченні.

З урахуванням множини позицій $tmPos$ множини представлення терміну tm, запис буде мати наступний вигляд:

$$r = < tm, lsn, nf, q, tmPos >, \quad (4)$$

де lsn – список опорних слів (іменників), які входять в термін в нормалізованому вигляді; nf – нормалізоване представлення терміну; q – кількість входження терміну в документ.

Для одного з варіантів представлення, записи в словнику термінів будуть мати наступний вигляд:

$$r = < t, abb, nf, tPos, f >, \quad (5)$$

де t – елемент множини tm, варіантів представлення терміну (можливо, після редагування експертом); abb – абревіатура терміну t, якщо вона використовувалась в документі; $tPos$ – індекс терміну t в документі; f – частота появи терміну в документі.

Частота f визначається за формулою:

$$f = \frac{q}{\sum_{i=1}^{nt} q_i}, \quad (6)$$

де nt – кількість термінів, виявлених в документі.

Кожен запис r в словнику предметної області R зберігається в постійній пам'яті. Таким чином, є можливість визначити спільний об'єм словника предметної області $V(R)$. Довжина вхідного рядку визначається методом *LEN* (довжина вхідного тексту в символах):

$$V(R) = (LEN(t) + LEN(add) + LEN(nf) + LEN(tPos) + LEN(f)) * n, \quad (7)$$

де n – кількість записів в словнику.

Для швидкого пошуку, індекс завантажується в оперативну пам'ять. Для реалізації пошуку, по відомих позиціях, не потрібно використовувати всю інформацію із словника R. Для цього на момент роботи, формується множина записів:

$$RI = \{ri_i\} i = 1, n.$$

Кожен запис має вигляд:

$$ri = < t, tPos > .$$

Розмір індексу в такому випадку буде мати вигляд:

$$V(RI) = (LEN(t) + LEN(tPos)) * n. \quad (8)$$

Для того, щоб здійснювати подальший пошук необхідного терміну, можна здійснити сортування таблиці термінів в алфавітному порядку:

$$R =>_{sort} (t, A- > Я). \quad (9)$$

Сортування запропоновано реалізувати методом «швидке сортування» [8]. Визначення необхідного терміну зі списку здійснюється логарифмічним (бінарним) пошуком. Після того як термін знайдено, можна отримати інформацію про його позицію в конкретному тексті. Всі метадокументи можна сортувати у відповідності з методом швидкого сортування по параметру ідентифікатора документу *DocId* (завчасне сортування). Використавши бінарний пошук, можна знайти потрібний документ. Після того як було отримано необхідний метадокумент, в ньому потрібно здійснити пошук вказаного номеру речення. Номер речення співпадає з номером рядка в метадокументі.

Кластеризація документів

Завчасне групування документів призначене для об'єднання маленьких документів у великі для коректного виділення термінів. Після виділення термінів необхідно розділити документи по змісту для підвищення швидкості пошуку необхідної інформації. Відомо багато способів кластеризації [9-12]. Групування документів, яке базується на кластеризації, а саме виявлення відстанні між документами, також розглянуто в роботі [13]. Для кластеризації, основаної на множинах термінів та їх частотах, можна використати метод, який описаний в роботі [13], оскільки, при обрахування відстані між документами, здійснюється повторне обрахування частот термінів, але не враховується кількість повторень терміну при визначені відстані між документами.

Представимо документ наступним чином:

$$D = \{< t, f >\}.$$

Нехай D_1 та D_2 два документи, які містять q_1 та q_2 термінів відповідно.

Відстань між двома документами представимо у вигляді функції:

$$d = \int (Nt, St, Qt), \quad (10)$$

де Nt – множина термінів, які не співпадають; St – рівень входження одного документу в інший (спільні терміни); Qt – відношення частот термінів, які співпадають для двох документів.

Нормалізовану відстань d_{12} між документами можна розглянути в трьох ситуаціях:

$(D_1 \in_t D_2) \vee (D_2 \in_t D_1)$ – документи співпадають (всі терміни з документу D_1 входять до документу D_2 або навпаки). В даному випадку відстань $d_{12} = 0$.

$D_1 \cap_t D_2 = \varnothing$ – документи відносяться до різних кластерів. Відстань між кластерами $d_{12} = 1$.

$(D_1 \setminus_t D_2) \cup (D_2 \setminus_t D_1) \neq \varnothing$ – існують терміни, які не є спільними для документів D_1 та D_2.

Множина термінів, які не співпадають Nt буде мати наступний вигляд:

$$Nt = (D_1 \setminus_t D_2) \cup (D_2 \setminus_t D_1).$$

Степінь входження одного документу в інший можна представити наступним чином:

$$St = D_1 \cap_t D_2.$$

Відношення частот термінів, які співпали Qt для документів, які порівнюються, буде мати вигляд:

$$Qt = (St \neq \varnothing \mid ((Qt_1 + Qt_2)/2),$$

де Qt_1 та Qt_2 – відношення кількості повторень спільних термінів до всієї кількості термінів в документі.

Таким чином, Qt_1 та Qt_2 можна представити наступним чином (ns – кількість спільних унікальних термінів для документів D_1 та D_2):

$$Qt_1 = \sum_{i=1}^{ns} \frac{f_{1i}(t_i) \in D_1}{q_1}; \quad (10)$$

$$Qt_2 = \sum_{i=1}^{ns} \frac{f_{2i}(t_i) \in D_2}{q_2}. \quad (11)$$

Отримані оцінки відстані між документами, на основі співставлення множини термінів та їхніх частот, дозволило застосувати один з відомих методів кластеризації [9-13]. На основі оцінки «простота/якість», було використано алгоритм агломеративної кластеризації [10].

Апробація рішень

Для апробації запропонованих рішень було створено програмний модуль TermsIndexing. Даний модуль виконує індексацію термінів і працює сумісно з програмою побудови словника предметної області [13]. Також представлений програмний модуль здійснює кластеризацію документів по описаних методах. Схема взаємодії системи побудови словника предметної області з програмним модулем TermsIndexing представлена на рисунку 1.

Для проведення дослідів запропонованого методу були використані тексти з різних областей науки та техніки. Досліди проводилися на 40 документах, кожен з яких містив в середньому 5500

слів, або 290 речень відповідно. Інтенсивність зустрічі терміну з найбільшою кількістю появ в середньому склав 60 повторень. Всього було виділено 21574 термінів з усіх документів. При застосуванні індексації, середній час виділення термінів з документів збільшився на 18%.

Досліди показали, що додання індексації в технологію побудови словника предметної області збільшило загальний час його побудови на 27%, що можна вважати допустимим. Об'єм словника, після додання індексації, збільшився на 35%.

Якість індексації було перевірено на 50 термінах. Для них усіх були знайдені документи та речення в документах, в які вони входили.

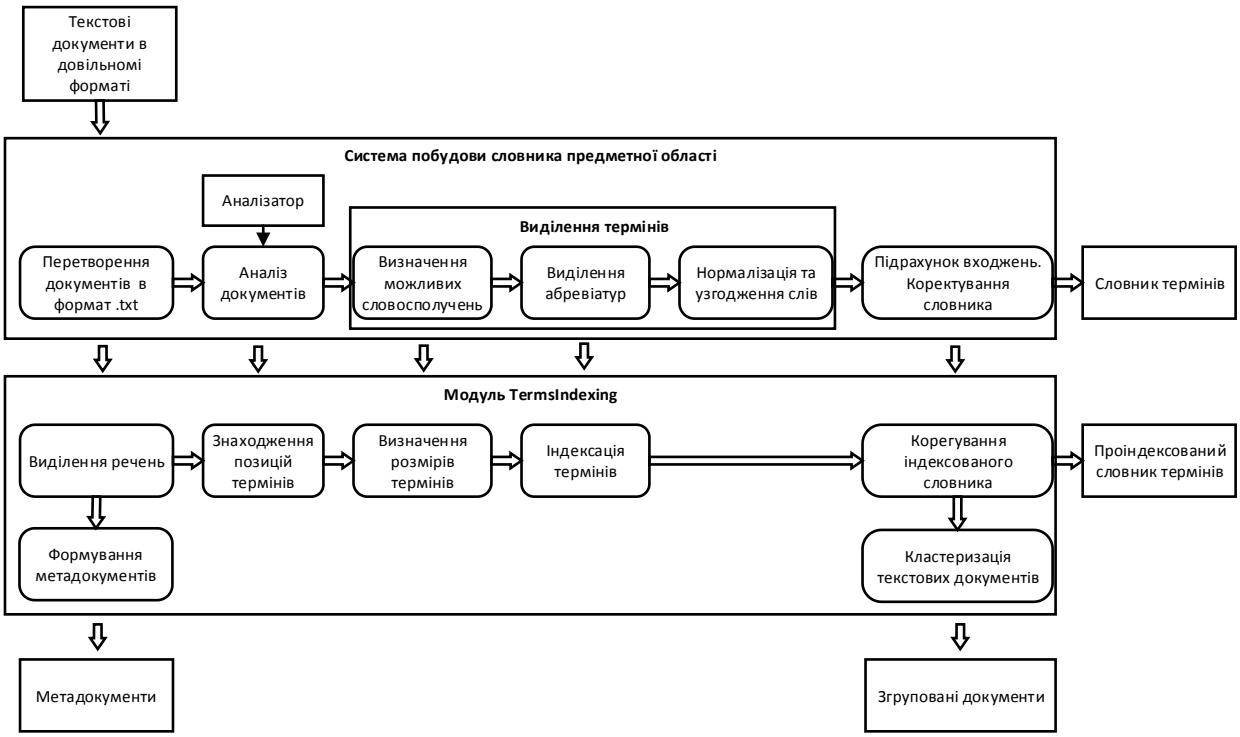

Рисунок 1 – Функціональна схема взаємодії програмного модуля TermsIndexing з системою побудови словника предметної області.

Висновки

Розроблено метод формування індексів виявлених термінів в документах корпоративних систем. Описано спосіб представлення позицій термінів та процес внесення даної інформації до словника предметної області. Представлено опис визначення розміру індексу та рекомендації правильному використанню індексу для ефективної роботи системи. Розроблено метод кластеризації документів з урахуванням наявності в них спільних термінів. Представлено алгоритм роботи системи. Реалізовано програмне рішення, яке дозволяє практично застосувати представлені методи в корпоративних системах.

Список літератури

1. Кунгурцев О. Б. Побудова словника предметної області на основі автоматизованого аналізу текстів українською мовою / О. Б. Кунгурцев, С. В. Ковальчук, Я. В. Поточняк, М. В. Широкоступ // Технічні науки та технології. – №3 (5), 2016. – С. 164-174.

2. Кунгурцев О. Б. Метод формування фрагментів тексту на основі розподілу термінів по документу / О. Б. Кунгурцев, С. В. Ковальчук // Електротехнічні та комп'ютерні системи. – 2017. – № 26 (102). – С.48–59.

3. Karmakar S. New Concept based Indexing Technique for Search Engine / S. Karmakar, S. Swarnakar // Indian Journal of Science and Technology, Vol 10(18), DOI: 10.17485/ijst/2017/v10i18/114018, May 2017.

4. Трифонов А. А. Алгоритмы построения инвертированного индекса для коллекции текстовых данных / А. А. Трифонов // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2013. – № 3 (27). – С. 52–61.

5. Büttcher S. Information Retrieval: Implementing and Evaluating Search Engines / Stefan Büttcher, Charles L. A. Clarke, and Gordon V. – MIT Press, 2010. – 606 c.

6. Bosch A. Estimating search engine index size variability: a 9-year longitudinal study / A. van den Bosch, T. Bogers, M. de Kunder // Scientometrics, 2016, Volume 107, Issue 2, pp 839–856.

7. Мизернов И. Ю. Анализ методов оценки сложности текста / Мизернов И. Ю., Гращенко Л. А. // Новые информационные технологии в автоматизированных системах, 18, С. 572—581.

8. Saadati R. Quicksort algorithm: Application of a fixed point theorem in intuitionistic fuzzy quasi-metric spaces at a domain of words / R. Saadati, S. M. Vaezpour, Y. J. Cho // Journal of Computational and Applied Mathematics, Vol. - 228, Issue 1, June 2009, Pages 219-225.

9. Novokhatska K. Application of Clustering Algorithm CLOPE to the Query Grouping Problem in the Field of Materialized View Maintenance / K. Novokhatska, O. Kungurtsev // Journal of Computing and Information Technology Vol 24,No 1 (2016) – P. 79 – 89.

10. Kogan J. Clustering Large and High-Dimensional Data / J. Kogan // University of Maryland, Baltimore, 2007.

11. Jacques J. Model-based clustering for multivariate functional data / J. Jacques, C. Preda // Computational Statistics & Data Analysis, Vol. - 71, March 2014, Pages 92-106.

12. Krishnasamy G. A hybrid approach for data clustering based on modified cohort intelligence and K-means / G. Krishnasamy, A. J. Kulkarnib, R. Paramesrana // Expert Systems with Applications, Vol. - 41, Issue 13, October 2014.

13. Кунгурцев А. Б. Метод автоматизированного построения толкового словаря предметной области / А. Б. Кунгурцев, Я. В. Поточняк, Д. Ф. Силяев // Технологический аудит и резервы производства — № 2/2(22), 2015. – С58 – 63.

References

1. Kungurtsev A.B., Kovalchuk S.V., Potochniak I.V., Shirokostup M.V. (2016). *Pobudova slovnyka predmetnoyi oblasti na osnovi avtomatyzovanogo analizu tekstiv ukrayinskoyu movoyu* [Creating the domain vocabulary on basis automated analysis of Ukrainian texts]. *Texnichni nauky ta texnologiyi* – Technical sciences and technology, no. 3(5), pp. 164 – 174 (in Ukrainian).

2. Kungurtsev A.B., Kovalchuk S.V. (2017). *Metod formuvannya fragmentiv tekstu na osnovi rozpodilu terminiv po dokumentu* [Formation method of text fragments based on the distribution of terms by document]. *Elektrotexnichni ta komp'yuterni systemy* – Electrotechnic and computer systems, – № 26 (102). – pp.48–59 (in Ukrainian).

3. Karmakar S., Swarnakar S. (2017). New Concept based Indexing Technique for Search Engine / Indian Journal of Science and Technology, Vol 10(18), DOI: 10.17485/ijst/2017/v10i18/114018 (In English).

4. Trifonov A.A. (2013). *Algoritmyi postroeniya invertirovannogo indeksa dlya kollektsii tekstovyih dannyih* [Algorithms for constructing an inverted index for a collection of text data]. *Izvestiya vyisshih uchebnyih zavedeniy. Povolzhskiy region. Tehnicheskie nauki* - Proceedings of higher educational institutions. Volga region. Technical science. – № 3 (27). – pp. 52–61 (In Russian).

5. Büttcher S.,Clarke C., Gordon V. (2010). Information Retrieval: Implementing and Evaluating Search Engines / MIT Press, – 606 p. (In English).

6. Bosch A., Bogers T., Kunder M. (2016). Estimating search engine index size variability: a 9-year longitudinal study / Scientometrics, Volume 107, Issue 2, pp 839–856 (In English).

7. Mizernov I., Hrashchenko L. (2018). *Analiz metodov otsenki slozhnosti teksta* [Analysis of text complexity assessment methods]. *Novyie informatsionnyie tehnologii v avtomatizirovannyih sistemah* - New information technologies in automated systems, pp. 572—581 (In Russian).

8. Saadati R., Vaezpour S., Cho Y. (2009). Quicksort algorithm: Application of a fixed point theorem in intuitionistic fuzzy quasi-metric spaces at a domain of words / Journal of Computational and Applied Mathematics, Vol. - 228, Issue 1, pp. 219-225 (In English).

9. Novokhatska K., Kungurtsev O. (2016). Application of Clustering Algorithm CLOPE to the Query Grouping Problem in the Field of Materialized View Maintenance / Journal of Computing and Information Technology Vol 24,No 1 – P. 79 – 89 (In English).

10. Kogan J. (2007). Clustering Large and High-Dimensional Data / University of Maryland, Baltimor (In English).

11. Jacques J., Preda C. (2014). Model-based clustering for multivariate functional data / Computational Statistics & Data Analysis, Vol. - 71, pp. 92-106 (In English).

12. Krishnasamy G., Kulkarnib A., Paramesrana R. (2014). A hybrid approach for data clustering based on modified cohort intelligence and K-means / Expert Systems with Applications, Vol. - 41, Issue 13 (In English).

13. Kungurtsev A.B., Potochniak I.V., Silyaev D.F. (2015). *Metod avtomatizirovannogo postroeniya tolkovogo slovarya predmetnoy oblasti* [The method of automated construction of the explanatory dictionary of the subject area]. *Tehnologicheskiy audit i rezervyi proizvodstva* - Technology audit and production reserves, № 2/2(22), pp. 58 – 63 (In Russian).

Надійшла до редакції 19.10.2018

С.В. КОВАЛЬЧУК
Одесский национальный политехнический университет, г. Одесса, Украина

УСОВЕРШЕНСТВОВАНИЕ МЕТОДА ИНДЕКСАЦИИ ТЕРМИНОВ В СГРУППИРОВАННЫХ ДОКУМЕНТАХ ДЛЯ РЕАЛИЗАЦИИ ПОИСКА В КОРПОРАТИВНЫХ СИСТЕМАХ

Разработан метод индексации терминов, который совмещен с процессом их выделения. Предложено формировать индекс из четырех компонентов: идентификатора документа, номера предложения в документе, позиции начала термина в предложении и количества слов в термине. Индексация позволяет выполнить в соответствии с запросом пользователя не только поиск документа, но и его фрагмента.

Ключевые слова: термин, корпоративная система, индексация, инвертированный индекс, словарь предметной области.

S. KOVALCHUK
Odessa National Polytechnic University, Odessa, Ukraine

IMPROVEMENT OF THE METHOD OF INDEXING TERMS IN GROUPED DOCUMENTS FOR THE REALIZATION OF SEARCH IN CORPORATE SYSTEMS

Finding relevant user information is the main goal of any information system. The process of allocation of terms is related to the sequential analysis of all sentences and words of the base text. There is a need to combine this process with the definition of the location of the term, it means enter indexation of terms.

The problem of quick and qualitative search of various terms in documents is related to the possibility of storing information about the location of the term, taking into account the number of words in it. The problem with unifying the procedure for the allocation of terms and their further indexation remains unresolved.

Attempts to index the terms during their selection in the documents to the subject vocabulary have been described. The process of selecting fragments of the text has also been described. But it was not described how to correctly form and place the index in the subject domain dictionary.

Despite of the fact that a lot of research has been done on this topic, they do not completely solve the problem of indexation of terms in corporate systems.

Aim of the research is the reducing the time to search for the necessary information by indexing the terms in corporate systems.

In the research process we developed the method of indexing of terms, associated with the process of allocation of terms. Also it was proposed to form an index of four components: the document ID, the sentence number in the document, the beginning of the term in the sentence and the number of words in the term. Indexing allows searching the document and snippets in documents. The method of document grouping process is described.

The method of forming the indexes of detected terms in the corporate system documents was developed. The method of recording positions of the terms and the process of entering this information into the domain knowledge dictionary are described. The method of determining the size of the index is described. Recommendations for the correct use of the index for the effective operation of the system are provided. The method of clustering of documents was developed taking into account the availability of common terms in documents. Software module called " TermsIndexing " was implemented.

Keywords: term, corporate system, indexing, inverted index, domain knowledge dictionary.