УДК 004.421

д.т.н. Самохвалов Ю.Я. (ВІТІ НТУУ„КПІ”)
Єрмоленко В.М.(ВІТІ НТУУ„КПІ”)

АНАЛІЗ МОДЕЛЕЙ ПОБУДОВИ СИСТЕМ ПОВНОТЕКСТОВОГО ПОШУКУ ДОКУМЕНТІВ У СИСТЕМІ ЕЛЕКТРОННОГО ДОКУМЕНТООБІГУ ЗБРОЙНИХ СИЛ УКРАЇНИ

У статті розкриті недоліки системи електронного документообігу Збройних Сил України та визначені шляхи її вдосконалення. Розглянуто основні моделі повнотекстового пошуку документів та сформовані вимоги до них. Обґрунтовано необхідність розробки нової моделі побудови підсистеми повнотекстового пошуку документів в системі електронного документообігу Збройних Сил України.

Самохвалов Ю. Я., Ермоленко В. М. Анализ моделей построения средств полнотекстового поиска в системе электронного документооборота Вооруженных Сил Украины. *В статье раскрыты недостатки системы электронного документооборота Вооруженных Сил Украины и определены пути их усовершенствования. Рассмотрены основные модели полнотекстового поиска документов и сформированы основные требования к ним. Обоснована необходимость разработки новой модели построения подсистемы полнотекстового поиска документов в системе электронного документооборота Вооруженных Сил Украины.*

U. Samokhvalov, V. Yermolenko The analysis of model building of the full-text means search in the electronic documents circulation system of the Armed Forces of Ukraine. *The article reveals weaknesses in the electronic documents circulation system of the Armed Forces of Ukraine and the ways of their improvement were detected. The basic models of a full-text search of documents were examined and the basic requirements for them were formed. It was grounded the necessity of a new model development of sub-text document search system in the electronic document circulation tsystem of the Armed Forces of Ukraine.*

Ключові слова: *система електронного документообігу, підсистема пошуку документів, повнотекстовий пошук, модель пошуку, індексація, нечіткий запит, логічний вираз.*

Актуальність. В даний час в Збройних Силах України особливу увагу направлено на створення єдиного інформаційного простору, що дозволить більш ефективно використовувати війська (сили) шляхом автоматизації роботи командирів, начальників і посадових осіб органів військового управління по управлінню підпорядкованими штабами, процесів приймання, обробки, відображення, документування та передачі інформації у повсякденній діяльності Збройних Сил України. Одним з основних напрямків створення єдиного інформаційного простору є розробка системи електронного документообігу Збройних Сил України (СЕДО).

Система електронного документообігу призначена для автоматизації процесу діловодства органів управління в Збройних Силах України. Основним її завданням є перехід від паперового життєвого циклу документів до електронного, надання електронним документам юридичної сили та забезпечення користувачів необхідними матеріалами при формуванні документів. У середньому на пошук необхідних відомостей користувач витрачає 30 – 40% часу в інформаційному масиві СЕДО, який включає в себе понад 10 млн. документів [1]. З огляду на це, ефективний пошук є однією з головних її функцій.

На даний момент підсистема пошуку СЕДО дозволяє здійснювати лише фактографічний пошук структурованих документів. Однак, практика використання цієї підсистеми показує, що у більшості випадків користувачам потрібен пошук документів без заздалегідь відомих атрибутів, тільки на підставі загальної теми, що неможливо зробити за допомогою існуючої підсистеми пошуку. Тому, розробка та впровадження моделей повнотекстового пошуку документів у СЕДО є актуальною задачею. Досвід використання інформаційно-пошукових систем показав [2], що для задоволення інформаційної потреби користувача систему необхідно спрямовувати на досягнення високих показників партинентності результатів пошуку – повноти та точності. З огляду на це, система повинна задовольняти наступним вимогам: використання у запитах логічних операторів, обробка

нечітких запитів, врахування невизначеностей природної мови (синоніми, мероніми, омоніми), здійснення семантичного пошуку, а також ранжування отриманих результатів.

Серед відомих моделей повнотекстового пошуку найбільш розповсюдженими є булева, векторна, розширена булева та ймовірнісна моделі, модель нечіткого пошуку, модель Mixed Min and Max та Paice-модель. Існують багато відомих інформаційно-пошукових систем, таких як Google, Yandex, Rambler, Meta, і т.п., які засновані на цих класичних моделях пошуку або їх комбінаціях. Але, по-перше, жодна з них повною мірою не задовольняє інформаційно-пошуковим вимогам, а по-друге, алгоритми їх роботи є комерційною таємницею.

Метою статті є аналіз моделей побудови підсистеми повнотекстового пошуку СЕДО з урахуванням основних інформаційно-пошукових вимог.

Булева модель [3, 4] є класичною моделлю інформаційного пошуку. Вона широко використовується в інформаційно-пошукових системах внаслідок простоти її реалізації. Ця модель дозволяє індексувати документи в масивах великого розміру та формувати пошукові запити за допомогою булевих операторів.

В рамках цієї моделі з множини документів $d^1, d^2, \dots, d^j, \dots, d^l$ формується документальний масив (колекція) D, а з термів $t_1, t_2, \dots, t_k, \dots, t_n$, які зустрічаються в D, формується словник T.

В булевій моделі кожний документ d^j описується вектором

$$d^j = (w_1^j, w_2^j, \dots, w_k^j, \dots, w_n^j), \quad (1)$$

де w_k^j – ваговий коефіцієнт терму t_k в документі d^j.

Цей коефіцієнт визначається як:

$$w_k^j = \begin{cases} 1, t_k \in d^j; \\ 0, t_k \notin d^j. \end{cases}$$

Запит користувача представляється логічним виразом, в якому терми t_p пов’язані між собою логічними операторами AND, OR та NOT. Далі такий вираз перетворюється у диз’юнктивну нормальну форму:

$$q = \bigvee_{i=1}^{r} q_{cc}^i,$$

де $q_{cc}^i = \bigvee_{p=1}^{m^i} t_p^i$ – і-та кон’юнктивна компонента запиту q, t_p^i – р-й терм в q_{cc}^i.

Міра близькості sim документа d^j до запита q обчислюється за виразом:

$$sim(d^j, q) = \begin{cases} 1, \text{якщо } \exists q_{cc}^i : (q_{cc}^i \in q) \wedge (\forall t_s^i, g_{t_s^i}(q_{cc}^i) = g_{t_s^i}(d^j)), \\ 0, \text{інакше.} \end{cases}$$

Тобто, $sim(d^j, q)=1$, якщо існує така кон’юнктивна компонента q_{cc}^i запиту, в якій інверсна функція $g_{t_s^i}(q_{cc}^i)$ кожного терма t_s^i даної компоненти співпадає з інверсною функцією $g_{t_s^i}(d^j)$ того ж терма t_s^i в документі d^j. У протилежному випадку $sim(d^j, q)=0$.

Векторна модель [4] є класичною алгебраїчною моделлю інформаційного пошуку. Більшість відомих інформаційних систем та систем класифікації інформації в тій чи іншій мірі засновані на використанні векторної моделі.

В цій моделі також формується колекція D та словник T. Документи та запити представляються векторами $d^j = (w_1^j, w_2^j, ..., w_k^j, ..., w_n^j)$, та $q^i = (w_1^i, w_2^i, ..., w_k^i, ..., w_n^i)$, у n-вимірному евклідовому просторі, де n – розмірність словника Т.

Для обчислення вагових коефіцієнтів w_k^j та w_k^i використовується TF-IDF метод [4], який є найефективнішим серед відомих [3-6]. Згідно [4] вагові коефіцієнти w_k^j та w_k^i обчислюються за формулою:

$$w_k^j = TF_k^j \cdot IDF_k, \; w_k^i = TF_k^i \cdot IDF_k,$$

де TF_k^j та TF_k^i – частота появи терміну t_k в документі d^j та запиті q^i відповідно, IDF_k – зворотна документальна частота терміну t_k для всієї колекції документів D.

Далі значення ваг w_k^j та w_k^i нормуються, що дозволяє розглядати документ та запити як ортонормовані вектори.

У векторній моделі можна використовувати різні методи обчислення міри близькості sim документа d^j до запита q^i [5, 6]. Разом з тим, використання скалярного добутку векторів дозволяє отримати більш достовірні показники повноти та точності результатів пошуку [6]:

$$sim(d^j, q^i) = \frac{\sum_{k=1}^{n} w_k^j \cdot w_k^i}{\sqrt{\sum_{k=1}^{n} (w_k^j)^2 \cdot \sum_{k=1}^{n} (w_k^i)^2}}. \qquad (2)$$

Латентно-семантичний аналіз (ЛСА) [7] є модифікацією векторної моделі пошуку. Цей аналіз використовується для визначення близькості значень слів та документів без використання зовнішніх баз знань. В основі ЛСА лежить гіпотеза про те, що між словами та їх контекстом, у якому вони використовуються, існують неявні(латентні) зв’язки.

Документи представляються векторами (1), з яких формується матриця C –“термін-на-документ”:

$$C = \begin{pmatrix} w_1^1 & w_1^2 & \dots & w_1^j & \dots & w_1^n \\ w_2^1 & w_2^2 & \dots & w_2^j & \dots & w_2^n \\ \dots & \dots & \dots & \dots & \dots & \dots \\ w_k^1 & w_k^2 & \dots & w_k^j & \dots & w_k^n \\ \dots & \dots & \dots & \dots & \dots & \dots \\ w_m^1 & w_m^2 & \dots & w_m^j & \dots & w_m^n \end{pmatrix},$$

де w_k^j – ваговий показник терма t_k^i в документі d^j.

Латентно-семантичний аналіз полягає в наступному:

1. Спочатку, проводиться сингулярне розкладання матриці C:

$$C = U У V^T,$$

де $U = CC^T$ – сингулярна матриця термів, $V^T = C^T C$ – сингулярна матриця документів, У – сингулярні числа матриці С.

2. Далі обчислюється матриця C_k – яка є малоранговою апроксимацією матриці С:

$$C_k = U_k У_k V_k^T,$$

де k – ранг матриці C_k.

Матриця C_k відображає структуру асоціативних зв’язків, які присутні у матриці С, що дозволяє частково вирішити проблему синонімії та полісемії термів. Ранг k для кожної колекції документів визначається емпіричним шляхом, та знаходиться у діапазоні від 150 до 400.

Запити, також представляються векторами: $q^i = (w_1^i, w_2^i, ..., w_k^i, ..., w_n^i)$. Далі формується вектор Q^j, який представляється як псевдодокумент:

$$Q^j = U_k^T \Sigma_k^{-1} q^i,$$

де q^i – початковий вектор запиту.

Для обчислення міри близькості між запитом та документом, між двома документами та між двома термінами використовують косинусну міру (2).

У роботі [6] показано, що обчислювальні витрати, які повязані з сингулярним розкладанням матриці значні. Невідомо ні одного вдалого досліду, в якому було б оброблено більш ніж мільйон документів. Саме ця обставина стала основною перешкодою для широкого використання ЛСА. Але, даний аналіз можна розглядати як м’яку кластеризацію, якщо кожну розмірність редукційного простору інтерпретувати як кластер, а значення, яке документ приймає на координатній осі, – як часткову приналежність кластеру.

Розширена булева модель. Практика використання пошукових систем показує, що кращі показники пошуку забезпечують гібридні моделі. В цих моделях процедура пошуку виконується у відповідності з булевою моделлю, а результати пошуку ранжуються по вагових коефіцієнтах у відповідності з векторною моделлю. Однією з таких гібридних моделей є розширена булева модель [8]. Вона надає можливість призначати вагові коефіцієнти термам, здійснювати пошук з частковою відповідністю та ранжувати його результати.

В рамках даної моделі документи описуються векторами (1), а запити – логічними виразами, які поділяються на диз’юнктивні

$$q_{or} = t_1 \vee^p t_2 \vee^p ... \vee^p t_S \vee^p ... \vee^p t_m,$$

та кон’юнктивні

$$q_{and} = t_1 \wedge^p t_2 \wedge^p ... \wedge^p t_S \wedge^p ... \wedge^p t_m,$$

де $p \in [1; \infty]$ – параметр, значення якого визначається під час формування запиту.

У роботі [13] показано, що для досягнення оптимальних показників повноти та точності пошуку параметр p повинний знаходитися в інтервалі [2; 5].

Міра близькості документа d^j до диз’юнктивного запиту обчислюються як:

$$sim(d^j, q_{or}) = \left(\frac{(w_1^j)^p + (w_2^j)^p + ... + (w_m^j)^p}{m} \right)^{\frac{1}{p}},$$

а до кон’юнктивного запиту – як:

$$sim(d^j, q_{and}) = \left(\frac{(1-w_1^j)^p + (1-w_2^j)^p + ... + (1-w_m^j)^p}{m} \right)^{\frac{1}{p}}.$$

Міра близькості документа d^j до більш загальних запитів (наприклад, $q = (t_1 \wedge^p t_2) \vee^p t_3$) обчислюється як:

$$sim(d^j, q)=\left(\frac{\left(1-\left(\frac{\left(1-w_1^j\right)^p+\left(1-w_2^j\right)^p}{2}\right)^p\right)^p+w_3^{jp}}{2}\right)^{\frac{1}{p}}.$$

Ця процедура може бути застосована рекурсивно, незалежно від кількості операторів AND/OR.

Ймовірнісна модель [6] пошуку базується на байєсовському підході. Характерною особливістю цієї моделі є оцінка ваг термів у документах шляхом обчислення ймовірностей присутності цих термів у релевантних та нерелевантних документах.

В рамках цієї моделі документи та запити представляються векторами

$$d^j = (w_1^j, w_2^j, ..., w_k^j, ..., w_n^j), \text{ та } q^i = (w_1^i, w_2^i, ..., w_k^i, ..., w_n^i),$$

де

$$w_k^j = \begin{cases} 1, t_k \in d^j \\ 0, t_k \notin d^j \end{cases}, w_k^i = \begin{cases} 1, t_k \in q^i \\ 0, t_k \notin q^i \end{cases}.$$

В якості оцінки міри близькості *sim* документа d^j до запита q^i використовуються ймовірності того, що користувач визнає документ релевантним або не релевантним запиту q^i. Ця міра обчислюється за наступним виразом:

$$sim(d^j, q^i) = \frac{P(R|d^j, q^i)}{P(\bar{R}|d^j, q^i)}, \quad (3)$$

де $P(R|d^j, q^i) = \frac{P(d^j|R, q^i)P(R|q^i)}{P(d^j|q^i)}$ та $P(\bar{R}|d^j, q^i) = \frac{P(d^j|\bar{R}, q^i)P(\bar{R}|q^i)}{P(d^j|q^i)}$ – ймовірності того, що документ d^j релевантний (R) та нерелевантний ($\bar{R}$) фіксованому запиту q^i відповідно; $P(d^j|R, q^i)$ та $P(d^j|\bar{R}, q^i)$ – ймовірності того, що якщо знайдений релевантний та нерелевантний документ, то його представлення має вигляд d^j; $P(R|q^i)$ та $P(\bar{R}|q^i)$ – апріорні ймовірності того, що за запитом q^i буде знайдений релевантний та нерелевантний документ; $P(d^j|q^i)$ – ймовірність того, що за запитом q^i буде знайдено документ d^j.

Зазначені ймовірності отримуються експертним шляхом на основі вибірки релевантних та нерелевантних документів.

В основі ймовірнісної моделі лежить припущенні про незалежність будь-якої пари термів в документі. Згідно цього припущення вираз (3) зводиться до:

$$sim(d^j, q^i) = \sum_{t_k \in d^j, q^i} \log \frac{rel_k(nrel - nrel_k)}{nrel_k(rel - rel_k)},$$

де rel_k та $nrel_k$ – відповідно кількість релевантних та нерелевантних документів колекції D, які містять термін t_k; *rel* та *nrel* – відповідно загальна кількість релевантних та нерелевантних документів запиту q^i.

Якщо для запиту q^i невідома кількість релевантних та нерелевантних документів, то для визначення міри близькості *sim* використовується алгоритм Okapi BM25 [6]:

$$sim(d^j, q^i) = \sum_{k=1}^{n} \log\left(\frac{N}{n(t_k)}\right) \times \frac{w_k^i \cdot w_k^j \cdot (a+1)}{w_k^i \cdot w_k^j + a \cdot \left(1 - b + b \cdot \frac{|d^j|}{avgdl}\right)},$$

де $|d^j|$ – довжина документа d^j (кількість в ньому слів); *avgdl* – середня довжина документа в колекції; *a*=2.0; *b*=0.75; *N* – загальна кількість документів в колекції *D*; $n(t_k)$ – кількість документів, які містять терм t_k.

Розглянута модель має високу математичну обґрунтованість, але має низку суттєвих недоліків. По-перше, для кожного нового запиту необхідно будувати нову вибірку релевантних та нерелевантних документів. По-друге, припущення про незалежність входження пари термів в документ значно обмежує пошукові можливості системи, т. я. воно повністю ігнорує контекст та семантику термів документа. З огляду на це ймовірнісна модель не отримала широкого використання в інформаційно-пошукових системах.

Модель нечіткого пошуку [3, 10] ґрунтується на теорії нечітких множин. Ця модель розширює класичну булеву модель можливістю ранжувати результати пошуку.

Для побудови інформаційно-пошукової системи на основі моделі нечіткого пошуку як і в попередніх моделях, використовується колекція *D* та словник *T*, а також формується словник лінгвістичних змінних:

$$L = (l^1, l^2, ..., l^i, ..., l^p).$$

Далі для кожної лінгвістичної змінної l^i формується терм-множина $A^i = (a_1^i, a_2^i, ..., a_k^i ..., a_r^i)$, універсальна множина U^i та функція приналежності $м_k^i$ елементів $u^i \in U^i$ до терма a_k^i.

В моделі нечіткого пошуку, як і в розширеній булевій моделі, запити поділяються на диз’юнктивні

$$q_{or} = t_1 \vee t_2 \vee ... \vee t_s \vee ... \vee t_m,$$

та кон’юнктивні

$$q_{and} = t_1 \wedge t_2 \wedge ... \wedge t_s \wedge ... \wedge t_m.$$

Кожен терм t_s цих запитів представляється як нечітка множина R_S, а документ d^j – як ступінь приналежності $м_{R_S}^j$ цій множині. Така ступінь задається наступним чином:

$$м_{R_S}^j = \begin{cases} 1, \text{ якщо } t_s \in d^j; \\ м_k^i, \text{ якщо } t_s = a_k^i \text{ та } u^i \in d^j; \\ 0, \text{ інакше.} \end{cases}$$

Міра близькості *sim* документа d^j до диз’юнктивного запиту обчислюється як:

$$sim(d^j, q_{or}) = \max(м_{R_1}^j, м_{R_2}^j, ..., м_{R_S}^j, ..., м_{R_m}^j),$$

а до кон’юктивного – як:

$$sim(d^j, q_{and}) = \min(м_{R_1}^j, м_{R_2}^j, ..., м_{R_S}^j, ..., м_{R_m}^j),$$

Ця модель дозволяє обробляти як чіткі так і нечіткі запити, що значно поширює функціональні можливості інформаційно-пошукових систем.

Модель MMM(Mixed Min and Max) [8] є розвитком моделі нечіткого пошуку. Документи та запити представляються як і в попередній моделі.

Міра близькості документа d^j до запиту q_{or} обчислюється як:

$$sim(d^j, q_{or}) = \text{ц}\cdot\max(\text{м}_{R_1}^j, \text{м}_{R_2}^j, ..., \text{м}_{R_S}^j, ..., \text{м}_{R_m}^j) + (1-\text{ц})\cdot\min(\text{м}_{R_1}^j, \text{м}_{R_2}^j, ..., \text{м}_{R_S}^j, ..., \text{м}_{R_m}^j),$$

а до запиту q_{and} – як:

$$sim(d^j, q_{and}) = \text{г}\cdot\min(\text{м}_{R_1}^j, \text{м}_{R_2}^j, ..., \text{м}_{R_S}^j, ..., \text{м}_{R_m}^j) + (1-\text{г})\cdot\max(\text{м}_{R_1}^j, \text{м}_{R_2}^j, ..., \text{м}_{R_S}^j, ..., \text{м}_{R_m}^j).$$

У роботі [9] показано, що використання лінійної комбінації максимумів та мінімумів у цих виразах при $\text{г}\in[0.5, 0.8]$ та $\text{ц}\in[0.2, 0.5]$ дає кращі показники повноти та точності пошуку, а також дозволяє більш ефективно ранжувати результати пошуку.

Раісе-модель [8] також є розвитком моделі нечіткого пошуку. Документи та запити представляються як в моделі нечіткого пошуку.

Міра близькості *sim* документа d^j до запитів q_{or} та q_{and} обчислюється як:

$$sim(d^j, q_{or}) = sim(d^j, q_{and}) = \frac{\sum_{s=1}^{m} r^{s-1} \text{м}_{R_S}^j}{\sum_{s=1}^{m} r^{s-1}}, \qquad (4)$$

де *r* – коефіцієнт, який для q_{or} запитів дорівнює 1, а для q_{and} запитів – 0.7.

Необхідно зазначити, що у (4) ступені приналежності $\text{м}_{R_S}^j$ для диз’юнктивних запитів використовуються у порядку спадання, а для кон’юнктивних – у порядку зростання.

Крім того в цій моделі, у відмінності від моделі МММ, при обчисленні міри близькості документів до запитів використовуються всі ступені приналежності $\text{м}_{R_S}^j$, що дозволяє значно підвищити показники повноти та точності пошуку.

У таблиці 1 наведено порівняльну характеристику розглянутих моделей повнотекстового пошуку на відповідність інформаційно-пошуковим вимогам:

Таблиця 1

У таблиці 2 наведено порівняльну характеристику найпопулярніших систем пошуку [10] :

	Використання булевих операторів у запитах	Ранжування результатів пошуку	Обробка нечітких запитів	Семантичний пошук	Врахування синонімії, полісемії, меронімії
Булева	+	–	–	–	Не враховується
Векторна	ТА	+	–	–	
Розширена булева	ТА/АБО	+	–	–	
Ймовірнісна	ТА	+	–	–	
Нечіткого пошуку	ТА/АБО	+	+	–	
Mixed Min and Max	ТА/АБО	+	+	–	
Раісе-модель	ТА/АБО	+	+	–	
Латентно-семантичний аналіз	ТА	+	+	–	Частково

Таблиця 2

	Google	Yandex	Rambler
Використання булевих операторів	+	+	+
Ранжування результатів пошуку	+	+	+
Обробка нечітких запитів	–	–	–
Аналіз запитів, сформованих природною мовою	+	–	–
Врахування синонімії, полісемії, метонімії термів запитів	+	–	–
Семантичний пошук	–	–	–
Об’єм індексу (кількість документів в массиві)	25 млрд.	1 млрд.	0,7 млрд.
Повнота	4%	4%	1%
Точність	0.3	0.28	0.26

Таким чином, аналіз сучасних моделей повнотекстового пошуку показав, що жодна з них не відповідає в повному обсязі інформаційно-пошуковим вимогам, які висуваються до підсистеми пошуку СЕДО. Крім того аналіз існуючих систем пошуку, таких як, Google. Yandex, Rambler показав, що вони також не відповідають цим вимогам. Разом з тим слід зазначити, що серед проаналізованих моделей пошуку найбільш ефективною є векторна модель, але вона не дозволяє обробляти нечіткі запити та використовувати булеві оператори у запитах. Тому для досягнення максимальної партинентності результатів пошуку необхідна розробка гібридної моделі на основі векторної моделі та моделі нечіткого пошуку, алгоритм роботи якої буде представлений у наступних публікаціях.

ЛІТЕРАТУРА

1. Наказ МОУ “Про прийняття на озброєння ЗСУ та організацію впровадження програмно-технічних комплексів, комплексів засобів зв’язку стаціонарних телекомунікаційних вузлів та терміналів відеоконференцзв’язку автоматизованої системи управління повсякденною діяльністю ЗСУ “Дніпро” № 185 від 05.04.2011.

2. Романенко Р. В. Сетевой информационный поиск. Практическое пособие / Р. В. Романенко, Г. В. Никитина. – Спб: Изд. “ПРОФЕССИЯ”, 2005.

3. Baeza-Yates R. Modern information retrieval / R. Baeza-Yates, B. Ribeiro-Neto // ACM Press Books. Addison Wesley, 1999.

4. Ландэ Д. В. Интернетика: навигация в сложных сетях. Модели и алгоритмы / Д. В. Ландэ, А. А. Санарский, И. В. Безусов. – М.: Книжный дом „ЛИБРОКОМ”, 2009.

5. Salton G. Term-weighting approaches in automatic text retrieval. / G. Salton, C. Buckley // Information Processing & Management. – 1988. – № 5. – С. 513 – 523.

6. Маннинг К.Д. Введение в информационный поиск / Маннинг К. Д., Рагхаван П., Шютце Х.; пер. с англ. Д. А. Клюшина. – М: “ООО Вильямс ”, 2011.

7. Landauer, T. K. An introduction to latent semantic analysis / T. K. Landauer, P. Foltz, D. Laham // Discourse Processes. – 1998. – № 25. – С. 259 – 284.

8. Шарапов Р. В. Пути расширения булевой модели поиска / Р. В. Шарапов, Е. В. Шарапова // ИСиТ. – 2009. – № 6.

9. Gudivada V. N., Raghavan V. V. Information Retrieval on the World Wide Web / V. N. Gudivada, V. V. Raghavan. – NY:IEEE Internet Computing. – 1997.

10. Kowalski G. Information Retrieval Architecture and Algorithms/ Kowalski G. – NY: Springer Science+Business Media. – 2011.

11. https://sites.google.com/site/infovrmir/home/f/sravnenie-poiskovyh-sistem.