УДК 004.91 DOI: 10.31673/2412-4338.2020.014552

Мілованова М.В., Бондарчук А.П. *Державний університет телекомунікацій, м. Київ*

ОЦІНКА ЕФЕКТИВНОСТІ ІНФОРМАЦІЙНОГО ПОШУКУ

В умовах науково-технічного прогресу стоїть завдання розробки принципово нових методів обробки, зберігання та контролю інформації. Сукупність інформаційних потоків, засобів збору, обробки та управління даними становить інформаційну систему. Спонукальною причиною здійснення інформаційного пошуку є інформаційна потреба, виражена у формі інформаційного запиту. Так як вимоги до швидкості пошуку, актуальності інформації з кожним днем стають все вище, то збільшуються і вимоги до методів і алгоритмів пошуку інформації. Основна мета інформаційного пошуку - допомогти користувачеві знайти інформацію, в якій він зацікавлений. Система відбирає з усього наявного безлічі інформації підмножина, яке задовольняє користувача. Існує значна кількість методів і алгоритмів інформаційного пошуку, проте зростання обсягів даних вимагає постійного поліпшення існуючих методів і розробку якісно нових підходів.

Необхідність оцінки ефективності якості пошуку можна назвати одним з основних завдань, вирішення якого повинно здійснюватися на стадії розробки інформаційно-пошукової системи. Якість пошуку може бути гарантовано за допомогою введення оцінок ефективності використовуваних методів пошуку, визначення типів джерел інформації, розробці алгоритмів та методів пошуку які найбільш ефективні в інформаційно-пошуковій системі яка розробляється.

У роботі розкрито основні аспекти інформаційного пошуку. Розглянуто методи пошуку інформації. Також проведена класифікація методів вибору джерел інформації та основні припущення при розробці інформаційно-пошукової системи, перераховані сучасні алгоритми пошуку. Розглянуто поняття релевантності, типи релевантності в залежності від відповідності результатів пошуку і запитом. Також введені характеристики оцінки релевантності результатів пошуку.

Ключові слова: *ефективність пошуку, інформаційно-пошукова система, алгоритм пошуку, формальна релевантність, онтологічна релевантність, пертинентність.*

Milovanova M.V., Bondarchuk A.P. *State University of Telecommunications, Kyiv*

THE INFORMATION SEARCH PERFORMANCE EFFICIENCY EVALUATION

Information is becoming a strategic resource of a new high-tech society. In the conditions of scientific and technological progress, there is a task is to develop fundamentally new methods of processing, storage and control of information. Information system is a combination of information flows, means of information collection, processing and data management. The main behavioral incentive for the implementation of information retrieval is the information need, expressed in the form of an information request. Since the requirements for search speed, the relevance of information every day are becoming higher, so are the requirements for methods and algorithms for finding information are continuously growing as well. The main purpose of information retrieval is to help the user to find information in which he is interested. There are a significant number of methods and algorithms for information retrieval, but the growth of data volumes requires continuous improvement of existing methods and the development of new approaches.

The most important task while developing an information retrieval system is to evaluate a searching efficiency. A searching quality depends on the efficiency of the used searching methods and algorithms, determining types of information sources correctly and implementing an efficiency assessment system.

The paper covers the main aspects of information retrieval, searching methods,

classification of information sources, the list of modern searching algorithms. The definition of relevancy and its criteria are given along with relevancy grades based on ranking of how well results matches the query. Moreover, different estimates for evaluating the relevance of search results have been introduced in the paper.

Keywords: *search performance efficiency, information retrieval system, search algorithm, formal relevance, ontological relevance, pertinence.*

Милованова М.В., Бондарчук А.П. *Государственный университет телекоммуникаций, г. Киев*

ОЦЕНКА ЭФФЕКТИВНОСТИ ИНФОРМАЦИОННОГО ПОИСКА

В условиях научно-технического прогресса стоит задача разработки принципиально новых методов обработки, хранения и контроля информации. Совокупность информационных потоков, средств сбора, обработки и управления данными составляет информационную систему. Побудительной причиной осуществления информационного поиска является информационная потребность, выраженная в форме информационного запроса. Так как требования к скорости поиска, актуальности информации с каждым днем становятся все выше, то увеличиваются и требования к методам и алгоритмов поиска информации. Основная цель информационного поиска - помочь пользователю найти информацию, в которой он заинтересован. Существует значительное количество методов и алгоритмов информационного поиска, однако рост объемов данных требует постоянного улучшения существующих методов и разработку качественно новых подходов.

Необходимость оценки эффективности качества поиска можно назвать одной из основных задач, решение которой должно осуществляться на стадии разработки информационно-поисковой системы. Качество поиска может быть гарантировано посредством ведения оценок эффективности используемых методов поиска, определения типов источников информации, разработке алгоритмов и методов поиска, которые наиболее эффективны в разрабатываемой информационно-поисковой системе.

В работе рассмотрены основные принципы работы информационного поиска. Рассмотрены модели поиска информации. Также проведена классификация методов выбора источников информации и основные допущенные при разработке информационно-поисковой системы, перечисленного современные алгоритмы поиска. Рассмотрено понятие релевантности, типы релевантности в зависимости от соответствия результатов поиска и поискового запроса. Также введены характеристики оценки релевантности результатов поиска.

Ключевые слова: *эффективность поиска, информационно-поисковая система, алгоритм поиска, формальная релевантность, онтологическая релевантность, пертинентность.*

Вступ.

В даний час наявність потрібної інформації є важливим аспектом будь-якої діяльності людини. Далеко не завжди необхідна інформація знаходиться в легкому доступі, частіше доводиться займатися її пошуком в різноманітних джерелах.

Метою даної роботи є аналіз основних понять інформаційного пошуку, а також сучасних методів для його здійснення.

У реальній практиці особі, що приймає рішення, або досліднику доводиться працювати в умовах інформаційної невизначеності [1] або інформаційної асиметрії [2]. Іноді інформаційну асиметрію трактують як семантичний розрив. Але і в цьому випадку інструментом подолання всіх перерахованих факторів є інформаційний пошук (Information Retrieval, IR) [3]. У процесі інформаційного пошуку проводять якісне і смислове оцінювання інформаційного об'єкта. Тобто по суті інформаційний пошук є набором методів когнітивного аналізу [4]. Інформаційний пошук є однією їх технологій збору інформації, необхідної для підтримки прийняття рішень. Інформаційний пошук застосовується для задоволення інформаційної потреби, продиктованої інформаційної ситуацією [5].

Тема інформаційного пошуку є популярною в сфері науково-дослідницької діяльності. Розглянемо деякі праці з цієї тематики. В. Я. Цветков в своїй роботі визначили що таке інформаційна невизначеність та визначеність в науках про інформацію [1], а також представив інформаційні моделі об'єктів, процесів і ситуацій [6]. У статті І.М. Загідулін була

розглянута нова система з пошуку інформації, заснована на контексті даних [7]. Продемонстрували особливості побудови пошукового запиту на основі нечітких множин - Р.В. Шарапов і Є.В. Шарапова [8].

З точки зору використання комп'ютерної техніки «інформаційний пошук» - сукупність логічних і технічних операцій, що мають кінцевою метою знаходження документів, відомостей про них, фактів, даних, релевантних запиту споживача. «Релевантність» – встановлюється при інформаційному пошуку відповідність змісту документа інформаційного запиту або пошукового образу документа пошуковому припису. Системи, що забезпечують реалізацію подібного пошуку інформації, називаються пошуковими системами (ПС). ПС з великим набором функцій і можливостей зазвичай входять до складу СУБД і іменуються інформаційно-пошуковими системами (ІПС). *Інформаційно-пошукова система* трактується і як система, що забезпечує пошук і відбір необхідних даних на основі інформаційно-пошукової мови і відповідних правил пошуку, а база даних - як сукупність засобів і методів опису, зберігання і маніпулювання даними, що полегшують збір, накопичення та обробку великих інформаційних масивів. Організація різних БД відрізняється видом об'єктів даних і відносин між ними.

Функціонування сучасних ІПС базується на двох припущеннях:

1) документи, необхідні користувачеві, об'єднані наявністю певної ознаки або комбінації ознак;

2) користувач здатний вказати цю ознаку.

Вибір джерела інформації безпосередньо залежить від сфери застосування тексту пошукового запиту. Відповідно до підходу до пошуку виділяють такі методи:

1) метод суцільного обстеження, що має на увазі під собою досконалий розгляд всіх джерел, що суттєво ускладнює процес пошуку;

2) метод вибіркового обстеження, де вивчаються дані певних джерел, що робить спосіб більш раціональним з точки зору людини;

3) метод інтуїтивного підходу, що підходить для досвідченого користувача, який володіє професійними навичками пошуку;

4) метод планового пошуку, де процес здійснюється по заздалегідь підготовленій моделі;

5) метод індуктивного пошуку, де факти розслідуються від суджень до узагальнень;

6) метод дедуктивного пошуку, де процес протилежний індуктивному.

В мережі Інтернет для зберігання інформації застосовуються сервери, де відбуваються процеси обробки і структурування, а так само реєстрації одержуваних даних, для спрощення пошуку за окремими словами або сферами застосування. Пошукові системи інформаційного типу є комплексами забезпечення даними всіх користувачів мережі Інтернет і вдають із себе так звані бази даних інформації. Такі системи можуть бути використані як одним користувачем, так і організацією різної сфери діяльності.

Предметом пошуку виступає інформаційна потреба користувача, неформально виражена в пошуковому запиті. І критерій пошуку, і його результати недетерміновані. Цими ознаками інформаційний пошук відрізняється від «пошуку даних», який оперує набором формально заданих предикатів, має справу зі структурованою інформацією і чий результат завжди детермінований. Теорія інформаційного пошуку вивчає всі складові процесу пошуку, а саме, попередню обробку тексту (індексування), обробку і виконання запиту, ранжування, призначений для користувача інтерфейс і зворотний зв'язок. [9].

Критерієм результату пошуку є отримання користувачем списку документів, одного документа або їх частин, максимально задовольняє його потребам, сформульованим в пошуковому запиті. В ІПС прийнято формувати список отриманих в результаті пошуку документів по їх релевантності. Розрізняють критерії смислового і формальної відповідності між пошуковим приписом і видаються документом.

Повнота і точність пошуку є взаємопов'язаними показниками. Збільшення одного з них веде до зниження іншого. В сучасних ІПС при збалансованому пошуку їх значення

становить приблизно 70%. Слід враховувати ситуацію, при якій список виданих пошуковою системою посилань містить кілька, а часом і десятки різних адрес з одним і тим же текстом. Подібні посилання характеризуються як дублікати. З них, при підрахунку коефіцієнтів враховується тільки один документ.

З огляду на, що ідеальний результат пошуку повинен задовольняти вимогам єдиності, повноти і несуперечності, отримуємо, що різні види пошуку визначають різні вимоги до функціональних можливостей системи в частині оцінювання результату. Однак, для випадку предметного пошуку доказ повноти є тривіальним: непорожній результат пошуку підтверджує факт існування (чи відсутності) об'єкта, що володіє шуканими властивостями. При цьому результат тематичного пошуку багатоаспектний і вимагає подальшої систематизації - ще одного процедурного кроку для впорядкування отриманого безлічі об'єктів за значеннями не визначений явно підстави. У свою чергу, проблемний пошук передбачає вже дворівневу систематизацію.

Алгоритми і методи пошуку. Відомі різні пошукові алгоритми. Прямим називається пошук, алгоритм якого ґрунтується на послідовному перегляді документів.

Прямий пошук - пошук безпосередньо за текстом документів [9], без попередньої обробки (без індексування). Прямий пошук тексту полягає в перегляді рядка (зліва направо) і послідовному порівняння кожної позиції з шуканим підрядком. Для цього порівнюють усі символи.

Інші алгоритми вимагають «індексування», попередньої обробки документів, при якій створюється допоміжний файл (індекс), покликаний спростити і прискорити пошук. Це алгоритми інвертованих файлів, суфіксних дерев, сигнатур.

Сигнатура (signature, підпис) - множина хеш-значений слів деякого блоку тексту. При пошуку по методу сигнатур усі сигнатури усіх блоків колекції видимі послідовно у пошуках збігів з хеш-значеннями слів запиту.

При інформаційному пошуку використовують поняття моделі пошуку. Модель пошуку – це деяке спрощення реальності, на підставі якої створюється оцінна формула, що дозволяє ІПС прийняти рішення: який документ вважати знайденим і як його ранжувати. Моделі інформаційного пошуку прийнято ділити на три види: теоретико-множинні (булева, нечітких множин, розширена булева); алгебра (векторна, узагальнена векторна, латентно-семантична, нейромережева) і імовірнісні.

Існують моделі пошуку, що аналізують сенс тексту, наприклад модель латентно-семантичного індексування (виявлення прихованого змісту). Ця модель алгебри грунтована на сингулярному розкладанні прямокутної матриці, що асоціює слова з документами. Елементом матриці є частотна характеристика, що відбиває міру зв'язку слова і документу [11].

Оцінка якості пошуку. Яка б не була модель, пошукова система потребує вдосконалення, яке зазвичай відбувається на підставі оцінки якості пошуку та за результататми налаштування параметрів. Оцінка якості – це ідея, фундаментальна для теорії пошуку. Завдяки оцінці якості можна говорити про застосовність або непридатність тієї або іншої моделі пошуку.

В результаті пошуку виявляється деяка інформація - *Ip*, яка може більшою чи меншою мірою задовольняти дослідника. Залежно від відповідності між метою пошуку *In* і результатом пошуку *Ip* можливі різні ситуації, що характеризуються поняттям релевантності (рис. 1). Можливі три типи відповідності між результатом пошуку і запитом: формальна релевантность, онтологічна релевантність, пертинентність [9, 12].

Формальною релевантністю називають відповідність пошукового образу пошуковому припису за формальними ознаками. За цими ознаками здійснюється відбір запитів і видача результатів пошуку. Формальна релевантність, як правило, далека від того, що хоче отримати дослідник. Онтологічна релевантність – відповідність пошукового образу пошуковому припису за смисловими ознаками. Вона передбачає порівняння запиту *In* і

результату пошуку *Ip* на семантичному (смисловому) рівні. Зокрема, при документальному пошуку порівняння відбувається на природній мові [12].

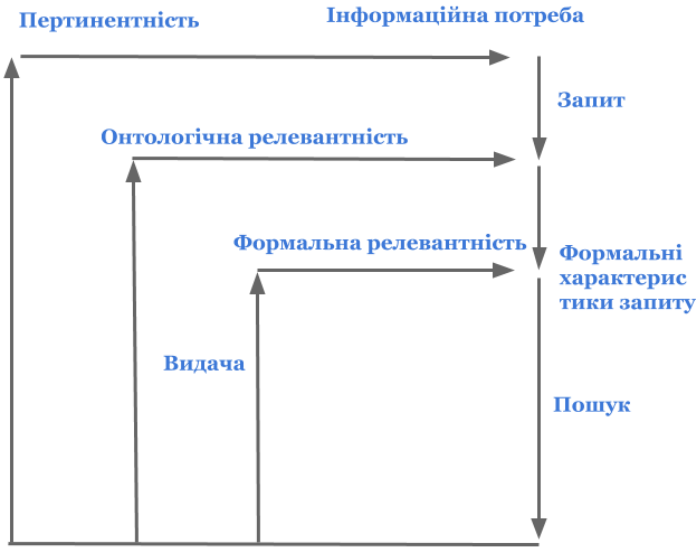

Рис. 1. Співвідношення між пертинентністю, онтологічною та формальною релевантністю

Інформація, що має смислову релевантність, формується на основі семантичної операції, тому вона несе в собі долю невизначеності. Критерій смислової відповідності формується людиною і встановлює відповідність між виданою системою інформацією і сенсом запиту. Це ставить перед дослідником додаткове завдання – точно визначити, релевантний або не релевантний результат пошуку. Таке завдання вирішується когнітивними методами [11]. Формальна і смислова релевантність можуть не задовольняти інформаційні потреби користувача. Найбільш повним критерієм відповідності результатів пошуку запиту вважається пертинентність. Пертинентністю називають повну відповідність знайдених знань або даних інформаційним потребам, яке встановлене при інформаційному пошуку.

Слід зазначити, що результат пошуку визначається не лише правильно побудованим запитом, але наявністю інформації про те, що необхідно шукати. Як правило, в результаті пошуку видається великий об'єм інформації, яка володіє формальною релевантністю. Ця інформація аналізується і, якщо необхідно, в ній проводиться уточнювальний пошук. Цей аналіз і додатковий пошук зменшують об’єм початково отриманих даних і створюють поле смислової релевантності. Таким чином, парадигматичний [16] ланцюжок: «початкові дані → формально релевантні дані → дані з онтологічною релевантністю → пертинентні дані» реалізується в технологічній послідовності: ІПС → оператор → експерт. Результат інформаційного пошуку доцільно оцінювати для того, щоб порівнювати різні пошукові технології і системи (рис. 2).

На рис. 2 використовуються наступні позначення:

– *T* - загальний час пошуку;
– *Vr = a b* - обсяг пошуку;
– *Vf = a da b db* - обсяг фонду, в якому виконаний пошук;
– *Vf = Vr dV* - обсяг фонду;
– *dV = da db* - частина обсягу фонду, не використана при пошуку;
– обсяг *a* – це обсяг формально релевантної інформації.
– обсяг *or* – це обсяг онтологічно релевантної інформації;
– обсяг *er* – це обсяг пертинентної інформації в результатах пошуку

Зведення цих характеристик наводиться в таблиці 1.

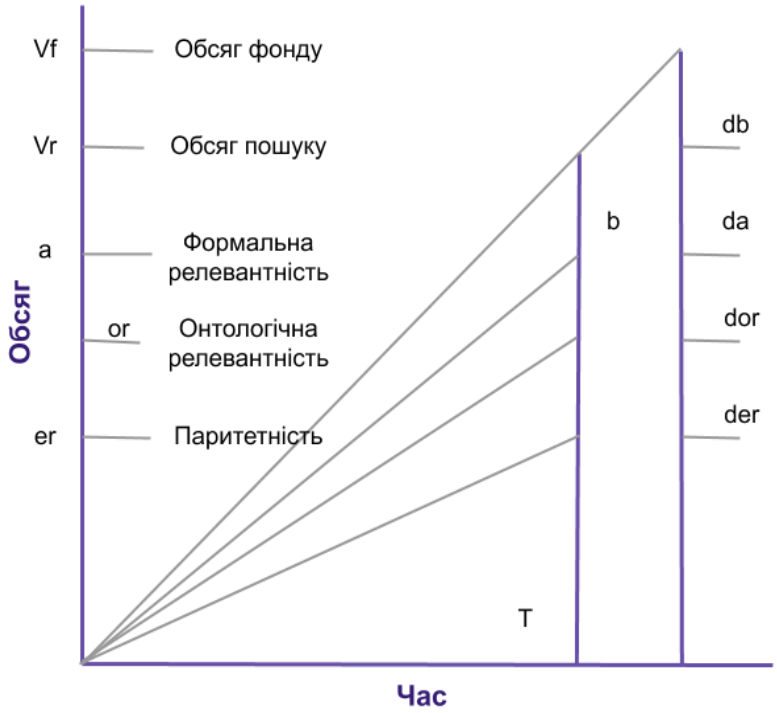

Рис. 2. Характеристики пошуку

Таблиця 1

Склад документів, наявних в інформаційному масиві і виданих в результаті пошуку

Видача	Формально релевантні	Онтологічно релевантні	Пертінентні	Нерелевантні	Всього
Видано	*a*			*b*	*a + b*
		or			
			er		
Не видано	*da*	*dor*	*der*	*db*	
Всього	*a + da*	*or + dor*	*er + der*	*b + db*	*a + b + da+ db*

Прирости *da; dor; der; d*b - залишки відповідних характеристик у фонді, які не потрапили в результат пошуку. Можна ввести наступні оцінки результатів інформаційного пошуку:

– повнота пошуку по формальній та онтологічній релевантности: $P = a/(a + da)$;

– повнота по пертинентності або коефіцієнт релевантності пошуку: $R = a/(a + da)$;

– коефіцієнт релевантності фонда: $Rf = (a + da)/Vf$;

– коефіцієнт онтологічності пошуку: $Ro = or/Vr$;

– коефіцієнт пертинентності пошуку: $Rp = er/Vr$;

– швидкість пошуку за час *T*: $VT = Vr/T$;

– ефективність пошуку за час *T*: $Et = Vr/Vf$.

Ці показники дають можливість оцінювати динамічні характеристики пошуку і якісно оцінювати результати пошуку. Введені характеристики дають можливість оцінювати ефективність різних ІПС і проводити порівняльний аналіз різних пошукових систем. Ефективність інформаційного пошуку вимірюється сукупністю різних показників, у тому числі технічною і економічною ефективністю. Технічна ефективність інформаційно-пошукової системи або технології визначається як міра виконання функції пошуку.

Економічна ефективність пошуку оцінюється за вартістю виконання цих функцій. Вартісні чинники можуть змінюватися з часом і регулюватися самим споживачем.

Висновок. Аналіз є необхідним компонентом інформаційного пошуку, оскільки на його основі приймається рішення про завершення або продовження пошуку. Існують спеціальні завдання інформаційного пошуку, рішення яких дозволяє розширювати процес пошуку. Інформаційний пошук може мати багатоаспектне представлення і не зводитися до простого перегляду і аналізу утримуваного якогось масиву.

Крім того, пошук як технологічна операція вимагає витрат. Чим більш тривалий пошук, тим більше вірогідності знаходження потрібної інформації. Але чим більш тривалий пошук, тим більше витрати. Конфлікт обумовлений протиріччям між витратами на пошук і необхідністю повного аналізу інформації. Цінність інформації при її пошуку може визначатися різними чинниками, у тому числі і витратами на її отримання, проте з часом ціна інформації, як товару може мінятися.

Сучасні технології інформаційного пошуку є комплексними технологіями, що включають інформаційне і когнітивне моделювання, а також застосування інформаційних одиниць. Сучасні технології інформаційного пошуку включають багаторівневу оцінку релевантності, що ускладнює оцінку результатів пошуку. Технології інформаційного пошуку вирішують завдання зняття інформаційної невизначеності, зменшення інформаційної асиметрії і подолання семантичного розриву. З цієї причини вони інтегровані в систему інформаційних технологій, пов'язану з обробкою інформації і управлінням.

Список використаної літератури

1. Цвєтков В. Я. Інформаційна невизначеність і визначеність в науках про інформацію. // Інформаційні технології. 2015. № 1. С. 3-7.

2. Васютинська С. Ю. Інформаційна асиметрія в освітніх технологіях. // Освітні ресурси і технології. 2016. № 4 (16). С. 14-20.

3. Іон А. Л. Алгоритми зменшення семантичних прогалин у системах пошуку зображень. // 2-я конференція 2009 року про взаємодію людини. IEEE, 2009, С. 97-102.

4. Цвєтков В. Я. Когнітивна наука пошуку інформації. // Європейський журнал психологічних досліджень. 2015. Vol. 5. Iss. 1. P. 37-44.

5. Ожерельева Т. А. Інформаційна ситуація як інструмент управління. // Слов'янський форум. 2016. № 4 (14). С. 176-181.

6. Цвєтков В. Я. Інформаційні моделі об'єктів, процесів і ситуацій. // Дистанційне і віртуальне навчання. 2014. № 5. С. 4-11.

7. Загідулін І.М. Використання контексту в інформаційному пошуку. // Системи управління та інформаційні технології. 2011. Т. 45. № 3.2. С. 297-301.

8. Шарапов Р.В., Шарапова Е.В. Застосування теорії нечітких множин в інформаційному пошуку. // Методи і пристрої передачі і обробки інформації. 2009. № 11. С. 427-432.

9. Романов В. П. Теоретичні основи інформатики. Інформаційні структури і фактографічний пошук інформації. М., РЕА ім. Г. В. Плеханова, 1996. 190 с.

10. Болбаков Р. Г. Розвиток і застосування когнітивно-семантичних методів і алгоритмів в мультимедійних освітніх портальних системах: Дис. ... канд. техн. наук. М .: МІРЕА, 2013. 184 с.

11. Цвєтков В. Я. Дескриптивні і прескриптивний інформаційні моделі. // Дистанційне і віртуальне навчання. 2015. № 7. С. 48-54.

12. Стоєва Д. Р. Систематизація інформаційних моделей. // Перспективи науки і освіти. 2015. № 4. С. 13-18

References

1. Tsvetkov V. Y. (2015) “Information certainty and uncertainty in information sciences”. Information technologies. № 1. P.3-7.

2. Vasyutinska S. Y. (2016) “Information asymmetry in educational technologies”. Educational resources and technologies.. № 4 (16). P.14-20.

3. Ion A. L. (2009) “Algorithms for reducing the semantic gap in image retrieval systems”. 2nd Conference on Human System Interactions. IEEE. P. 97-102.

4. Tsvetkov V. Y. (2015) “Cognitive Science of Information Retrieval”. European Journal of Psychological Studies.. Vol. 5. Iss. 1. P. 37-44.

5. Ozherelieva T. A. (2016) “Information situation as a management tool”. Slavic Forum. № 4 (14). P. 176-181.

6. Tsvetkov V. Y. (2014) “Information models of objects, processes and situations”. Distant and virtual learning. № 5. P. 4-11.

7. Zagidulin I. M. (2011) “Use of context in information search”. Management systems and information technologies. T. 45. № 3.2. P. 297-301.

8. Sharapov R. V., Sharapova E. V. (2009) “Application of fuzzy set theory in information retrieval”. Methods and devices for information transfer and processing. № 11. P. 427-432.

9. Romanov V. P. (1996) “Theoretical Foundations of Informatics. Information structures and factual search for information”. M. REA. 190 p.

10. Bolbakov R. G. (2013) “Development and “application of cognitive-semantic methods and algorithms in multimedia educational portal systems. P. 184.

11. Tsvetkov V. Y (2105) “Descriptive and prescriptive information models”. Remote and virtual training. № 7. P. 48-54.

12. Stoeva D. R. (2015) “Systematization of information models”. Prospects of science and education.. № 4. P. 13-18.