УДК 025.4.036

# МЕТОДИКА ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ТЕМАТИКО-ОРИЕНТИРОВАННОГО ИНТЕРНЕТ-ПОИСКА

**Сергеев А.Ю., Тютюнник В.М.**

*ФГОУ ВПО «Тамбовский государственный технический университет», Тамбов, e-mail: vmt@tmb.ru*

Разработана и протестирована методика повышения эффективности поиска, созданная на основе введённого ранее авторами показателя оценки эффективности интернет-поиска: глубина тематического охвата. Этот показатель оценивает зависимость тематической полноты найденной информации от количества просмотренных документов. Методика основана на выделении нижестоящих к поисковому дескрипторов по каждому из запросов, проведении отдельных поисковых сессий по каждому полученному поисковому образу запроса и интеграции результатов поисковых сессий для последовательного или параллельного просмотра в отношении искомых пертинентных документов. Экспериментально показано уменьшение минимального объёма поисковой выборки, в котором раскрыт семантический потенциал поискового термина при использовании предложенной методики оптимизации. Описаны четыре способа организации результирующей выборки. В результате применения предложенной методики удалось повысить эффективность тематико-ориентированного интернет-поиска более чем в три раза.

**Ключевые слова: информационный поиск, поисковая машина, эффективность поиска, оптимизация, точность, полнота, семантический потенциал, поисковая выборка, коэффициент семантического охвата, минимальный объём поисковой выборки**

# METHOD OF INCREASE OF SUBJECTS-FOCUSED INTERNET RETRIEVAL EFFICIENCY

**Sergeev A.Y., Tyutyunnik V.M.**

*TambovStateTechnologyUniversity, Tambov, e-mail: vmt@tmb.ru*

Is developed and tested thoroughly the procedure of an increase in the effectiveness of search, created on the basis of the introduced earlier by the authors index of the estimation of the effectiveness of the Internet-search: the thematic coverage level coefficient. This index evaluates the dependence of the subject completeness of the obtained information on a quantity of examined documents. Procedure is based on the isolation of the subordinate to the search descriptors on each of the demands, conducting of separate search sessions for each obtained search means of demand and integration of the results of search sessions for the sequential or parallel survey with respect to the desired pertinent documents. The decrease of the minimum size of search sample, in which is opened the semantic potential of search term with the use of the procedure of optimization proposed, had been experimentally shown. Four methods of organizing the resulting sample are described. As a result, the application of the procedure proposed it was possible to increase the effectiveness of the thematic-oriented Internet-search more than three times.

**Keywords: information retrieval, search engine, retrieval efficiency, optimization, precision, recall, semantic potential, search sample, thematic coverage level coefficient, search sample minimal volume**

В [1, 2] нами введён показатель оценки эффективности интернет-поиска – *глубина тематического охвата (thematic coverage level, TCL),* который показывает зависимость тематической полноты найденной информации от количества просмотренных документов, а также предложена новая методика оценки эффективности интернет-поиска, оперирующая семантической составляющей результатов поиска на основе коэффициента семантического потенциала поискового термина. Следующей задачей является разработка методики, позволяющей повысить эффективность тематико-ориентированного интернет-поиска с помощью минимизации объёма поисковой выборки, обеспечивающей тематическую полноту

Очевидно, что минимальный объем выборки, необходимый для обеспечения тематической полноты поиска $V_{min}$, при использовании поисковых терминов с показателем семантического потенциала $k = [1…3]$ является удовлетворительным. Как показано в [1, 2], работа с поисковой выборкой объёмом свыше пятнадцати документов не оправдана по причине резко снижающейся вероятности встретить новую информацию по теме поиска. Таким образом, тематико-ориентированный интернет-поиск с использованием поисковых терминов с показателем семантического потенциала $k > 3$ является низкоэффективным. Кроме того, в среднем в 40 % поисковых сессий при $k = [4…9]$ тематическая полнота не обеспечивалась в пределах поисковой выборки объёмом в 100 документов.

## Описание метода и результаты исследования

Разработка метода повышения эффективности поиска основывалась на следующих положениях:

1) новая информация по теме поиска сконцентрирована среди первых пятнадцати позиций поисковой выборки;

2) среднее значение оптимального объёма поисковой выборки представляет собой пятнадцать документов $S_{optim} = 15$;

3) объём поисковой выборки для обеспечения тематической полноты поиска при использовании поисковых терминов с показателем семантического потенциала $k > 3$ превышает оптимальный;

4) при использовании поисковых терминов с показателем семантического потенциала $k > 5$ вероятность получить полную информацию по теме составляет 40 %;

5) точность интернет-поиска представляет собой чаще всего константу $T \cong 0{,}56$ [3, 4];

6) включение оператора «ИЛИ» в запрос не имеет смысла при тематико-ориентированном поиске.

*Задача оптимизации поиска* сформулирована следующим образом: обеспечить такую тематическую полноту информации, получаемой в результате поисковой сессии (ПС) посредством уменьшения объёма поисковой выборки, чтобы семантический потенциал поискового термина был раскрыт максимально полно. Отсюда предлагаемый метод основывается на *гипотезе* о более высокой эффективности поиска при субституции поисковой сессии, содержащей поисковый термин $T_n$ с показателем семантического потенциала $k > 3$ на $k$ поисковых сессий по комплексу запросов $\sum_{1}^{k} T[k]_{n+1}$. Необходимость выделения отдельных поисковых сессий связана с положением 5.

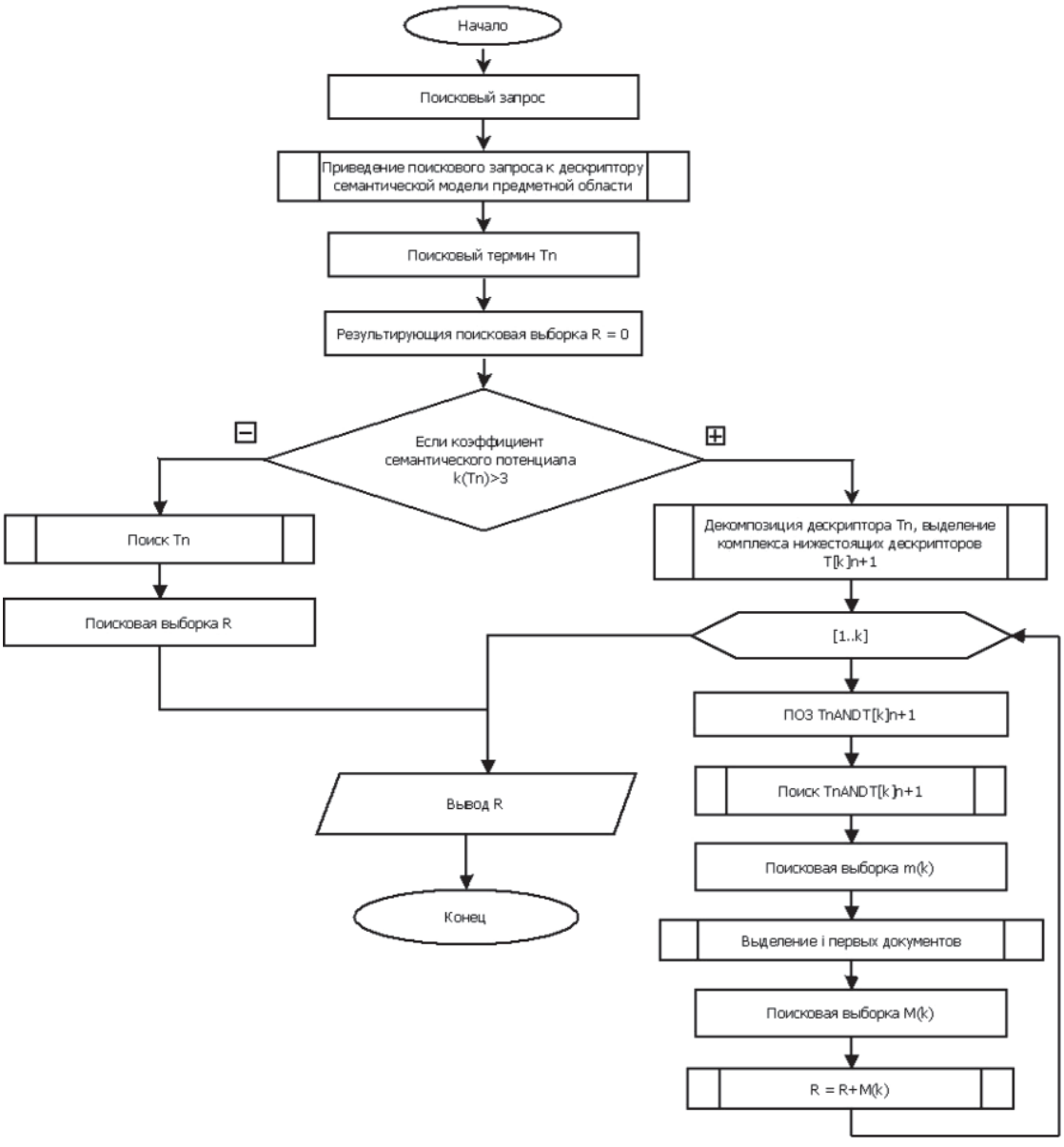


*Рис. 1. Алгоритм оптимизации объёма поисковой выборки согласно описанной методике*

*Методика* (алгоритм на рис. 1) основывается на выделении $k$ нижестоящих дескрипторов $T[k]_{n+1}$ по отношению к поисковому термину $T_n$, субституции исходной ПС на комплекс ПС по каждому из запросов $T[k]_{n+1}$ вида $\text{ПС}_0(T_n) \to \sum_1^k \text{ПС}_1\left(T_n ANDT[k]_{n+1}\right)$, проведении отдельных поисковых сессий по каждому полученному ПОЗ и интеграции результатов поисковых сессий для последовательного или параллельного просмотра в отношении поиска документов, пертинентных комплексу ПОЗ $\sum T[k]_{n+1}$. Параметр $i$, определяющий объём поисковой выборки для $\text{ПС}_1(T_n ANDT[k]_{n+1})$, установлен эмпирическим путём, его значение приведено ниже.

Для определения эффективности метода реализован эксперимент на комплексе исследованных ранее запросов с $k = [4…9]$. Фрагмент эмпирических данных отражён на рис. 2.

**Коэффициент семантического потенциала 5**

**Абстрактные типы данных**

***Объем поисковой выборки S(norm) = 86*** ***Объем поисковой выборки (поиск оптимизирован) S(opt) = 14***

| | | |
|---|---|---|
| Списки | Абстрактные типы данных Списки | 3 |
| Стеки | Абстрактные типы данных Стеки | 1 |
| Очереди | Абстрактные типы данных Очереди | 2 |
| Деки, Двухсторонние очереди | Абстрактные типы данных Деки, Двухсторонние очереди | 1 |
| Таблицы | Абстрактные типы данных Таблицы | 7 |

**Синтаксис программ**

***Объем поисковой выборки S(norm) = 34*** ***Объем поисковой выборки (поиск оптимизирован) S(opt) = 7***

| | | |
|---|---|---|
| Конкретный синтаксис | Синтаксис программ Конкретный синтаксис | 2 |
| Абстрактный синтаксис | Синтаксис программ Абстрактный синтаксис | 2 |
| Регулярное выражение | Синтаксис программ Регулярное выражение | 1 |
| Бесконтекстная (контекстно-свободная) грамматика | Синтаксис программ Бесконтекстная (контекстно-свободная) грамматика | 1 |
| Атрибутивная грамматика | Синтаксис программ Атрибутивная грамматика | 1 |

**Искусственный интеллект**

***Объем поисковой выборки S(norm) = 38*** ***Объем поисковой выборки (поиск оптимизирован) S(opt) = 13***

| | | |
|---|---|---|
| Модели когнитивных процессов | Искусственный интеллект Модели когнитивных процессов | 2 |
| Представление знаний | Искусственный интеллект Представление знаний | 2 |
| Рассуждение | Искусственный интеллект Рассуждение | 5 |
| Обучение | Искусственный интеллект Обучение | 1 |
| Прикладные системы искусственного интеллекта. Интеллектуальные системы, основанные на использовании знаний | Искусственный интеллект Прикладные системы искусственного интеллекта. Интеллектуальные системы, основанные на использовании знаний | 3 |

**Регистры процессоров**

***Объем поисковой выборки S(norm) = 9*** ***Объем поисковой выборки (поиск оптимизирован) S(opt) = 13***

| | | |
|---|---|---|
| Регистры общего назначения | Регистры процессоров Регистры общего назначения | 1 |
| Аккумуляторы. Накапливающие регистры | Регистры процессоров Аккумуляторы. Накапливающие регистры | 2 |
| Сдвиговые регистры | Регистры процессоров Сдвиговые регистры | 1 |
| Регистры чисел с плавающей запятой | Регистры процессоров Регистры чисел с плавающей запятой | 5 |
| Стековые регистры | Регистры процессоров Стековые регистры | 4 |

**Арифметические устройства**

***Объем поисковой выборки S(norm) = 69*** ***Объем поисковой выборки (поиск оптимизирован) S(opt) = 11***

| | | |
|---|---|---|
| Сумматоры. Полусумматоры | Арифметические устройства Сумматоры. Полусумматоры | 1 |
| Схемы образования дополнения | Арифметические устройства Схемы образования дополнения | 1 |
| Множительные устройства. Делительные устройства | Арифметические устройства Множительные устройства. Делительные устройства | 4 |
| Векторные арифметические устройства | Арифметические устройства Векторные арифметические устройства | 2 |
| Сдвигающие устройства. Сравнивающие устройства | Арифметические устройства Сдвигающие устройства. Сравнивающие устройства | 3 |

**Функционирование компьютерной памяти**

***Объем поисковой выборки S(norm) = 52*** ***Объем поисковой выборки (поиск оптимизирован) S(opt) = 12***

| | | |
|---|---|---|
| Чтение | Функционирование компьютерной памяти Чтение | 1 |
| Запись | Функционирование компьютерной памяти Запись | 1 |
| Доступ | Функционирование компьютерной памяти Доступ | 1 |
| Адресация | Функционирование компьютерной памяти Адресация | 2 |
| Поблочная передача | Функционирование компьютерной памяти Поблочная передача | 7 |

**Компьютеры общего назначения. Универсальные вычислительные машины**

***Объем поисковой выборки S(norm) = 19*** ***Объем поисковой выборки (поиск оптимизирован) S(opt) = 11***

| | | |
|---|---|---|
| Суперкомпьютеры. Супер-ЭВМ | Компьютеры общего назначения. Универсальные вычислительные машины Суперкомпьютеры. Супер-ЭВМ | 1 |
| Большие вычислительные машины. Универсальные ЭВМ. Мэйнфреймы | Компьютеры общего назначения. Универсальные вычислительные машины Большие вычислительные машины. Универсальные ЭВМ. Мэйнфреймы | 1 |
| Миникомпьютеры. Мини-ЭВМ | Компьютеры общего назначения. Универсальные вычислительные машины Миникомпьютеры. Мини-ЭВМ | 1 |
| Рабочие станции | Компьютеры общего назначения. Универсальные вычислительные машины Рабочие станции | 7 |
| Персональные компьютеры. ПЭВМ. Микрокомпьютеры. Микро-ЭВМ | Компьютеры общего назначения. Универсальные вычислительные машины Персональные компьютеры. ПЭВМ. Микрокомпьютеры. Микро-ЭВМ | 1 |

**S(norm)cp = 43,86** **S(opt)cp = 11,57**

**Стандартное отклонение (S(norm)) = 27,20** **Стандартное отклонение (S(norm)) = 2,30**

**S(norm) > 100, % = 0,00** **S(norm) > 100, % = 0,00**

*Рис. 2. Фрагмент эмпирических данных по различным запросам*

Эксперимент показал, что с использованием предложенного нами преобразования поисковой сессии при условии получения полной информации по теме поиска *среднее значение минимального объёма поисковой выборки, обеспечивающего тематическую полноту поиска* $V_{min}$, *составило 18 документов* (рис. 3), что в 3,7 раза меньше, чем значение, полученное при стандартном поиске (68 документов).

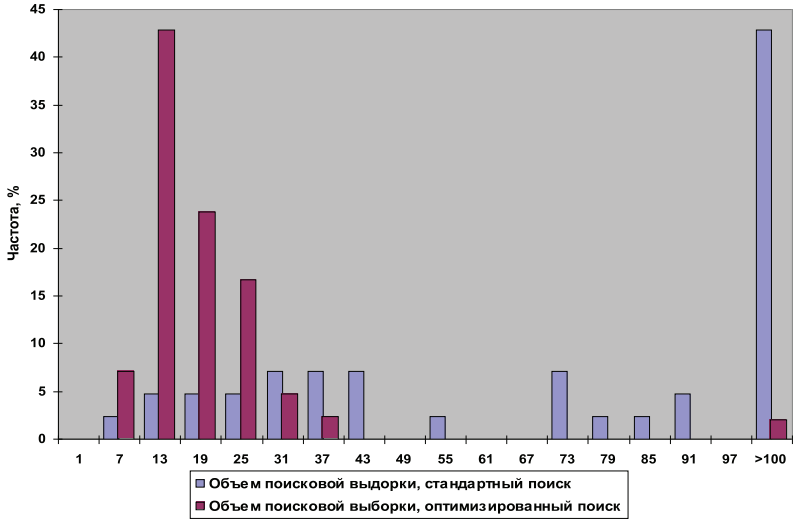


*Рис. 3. Объём поисковой выборки, раскрывающей семантический потенциал поискового термина, для простого и оптимизированного поиска*

Анализ эмпирических данных показал, что объём поисковой выборки лишь незначительно увеличивается с повышением значения коэффициента семантического потенциала поискового термина. Этот факт говорит об универсальности объёма поисковой выборки в 18 документов, т.е. предложенная методика раскрывает семантический потенциал поискового термина в среднем в объёме поисковой выборки, равном 18 документов (рис. 4).

Представленные данные получены без учёта перекрытия пертинентных документов среди промежуточных ПС. Фактически, документ, отражающий один семантический аспект исходного поискового термина $T_n$, может содержать также информацию о других аспектах искомой тематики. Другими словами, процесс получения комплекса искомой информации в реальности будет происходить быстрее. Для семи случайных ПС мы фиксировали *количество просмотров документов, необходимое для получения тематически полной информации*. Его среднее значение составило 12 документов.

Из распределения пертинентных страниц в поисковой выборке по каждому из запросов $T[k]_{n+1}$ (рис. 5) видно, что 97% пертинентных документов расположены в пределах первых семи позиций, 88% – четырёх.

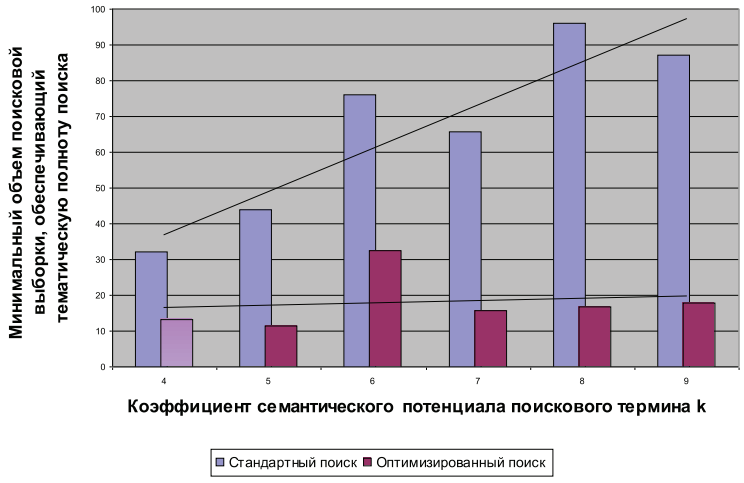


*Рис. 4. Зависимость минимального объёма поисковой выборки от коэффициента тематического потенциала поискового термина*

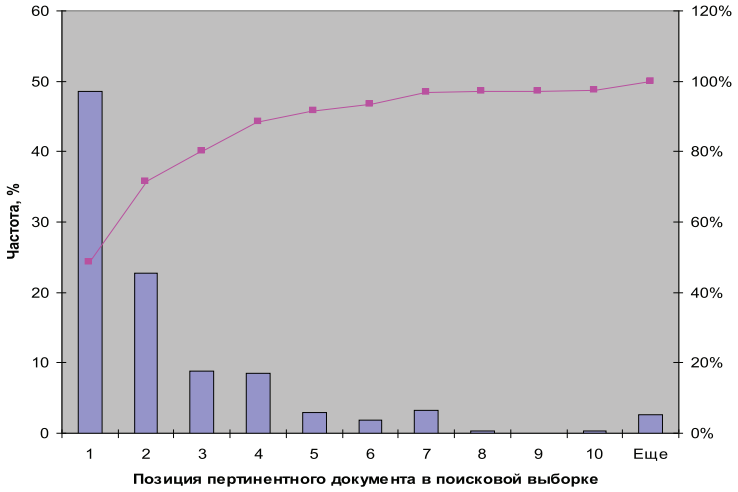


*Рис. 5. Распределение пертинентных страниц в промежуточных поисковых выборках*

Параметр $i$, определяющий объём промежуточной поисковой выборки для $ПС_1(T_nANDT[k]_{n+1})$, примем равным 7. Вероятность присутствия пертинентного документа в промежуточной поисковой выборке составит 97 %.

*Организация результирующей выборки.* При анализе результирующей поисковой выборки основными затратами пользователя, определяющими степень доступности результатов поиска и одновременно уровень их надёжности, являются: время просмотра; количество документов в результирующей выборке; количество поисковых терминов, которые необходимо держать в памяти; простой и логически понятный интерфейс. Приведём четыре варианта *максимально полного предоставления результатов оптимизированного поиска пользователю*, расположенные по снижению вероятных затрат среднестатистического пользователя:

1) последовательный просмотр поисковых выборок по каждому аспекту тематики поискового термина. Просмотр выборки продолжается до тех пор, пока не будет найден документ, содержащий информацию по данному аспекту тематики. *Документы, отражающие другие аспекты тематики поиска* ***игнорируются***. Пользователь оперирует одним поисковым термином. Объём поисковой выборки неограничен. Среднее значение минимального объёма, обеспечивающего тематическую полноту $V_{min}$ поиска, в данном случае составит 18 документов;

2) последовательный просмотр поисковых выборок по каждому аспекту тематики поискового термина. Просмотр выборки продолжается до тех пор, пока не будет найден документ, содержащий информацию по данному аспекту тематики. *Документы, отражающие другие аспекты тематики поиска* ***фиксируются***. Пользователь оперирует совокупностью поисковых терминов, раскрывающих семантический потенциал поискового термина. Объём поисковой выборки неограничен $V_{min} = 12$.

С целью минимизации объёма результирующей выборки [5] возможна реализация двух дополнительных вариантов:

1) интеграция результатов $ПС_1$, $i = 7$. Вероятность полного освещения тематики поиска 97 %. Объём поисковой выборки равен $4k$. Интерфейс подобен интерфейсу ПМ;

2) интеграция результатов $ПС_1$, $i = 4$. Вероятность полного освещения тематики поиска 88 %. Объём поисковой выборки равен $7k$.

## Выводы

1. Описана методика повышения эффективности поиска, разработанная в соответствии с полученными экспериментальными данными.

2. Экспериментально показано уменьшение минимального объёма поисковой выборки, в котором раскрыт семантический потенциал поискового термина при использовании предложенной методики оптимизации. Его среднее значение составило 18 документов, что в 3,7 раза меньше, чем значение, полученное для стандартного поиска (68 документов).

3. Объём поисковой выборки остаётся стабильным при повышении значения коэффициента семантического потенциала поискового термина (в противоположность п. 6).

4. Описаны четыре способа организации результирующей выборки. При исполь-

зовании одного из них среднее значение минимального объёма поисковой выборки, обеспечивающего тематическую полноту поиска, составило 12 документов, что в пять с лишним раз меньше аналогичного значения для стандартного поиска.

### Список литературы

1. Сергеев А.Ю., Тютюнник В.М. Эффективность тематико-ориентированного Интернет-поиска // Междунар. журн. эксперимент. образования. – 2012. – № 7. – С. 61–66.

2. Сергеев А.Ю., Тютюнник В.М. Методика оценки эффективности тематико-ориентированного Интернет-поиска с помощью минимизации объёма поисковой выборки // Фундамент. исследования. – 2013. – В печати.

3. Сергеев А.Ю., Тютюнник В.М. Разработка и тестирование методики оценки показателей эффективности сетевого информационного поиска // Формирование специалиста в условиях региона: Новые подходы: материалы 7 Всерос. межвузов. науч. конф., г. Тамбов, 5 марта 2008 г. – Тамбов; М.; СПб.; Баку; Вена: Изд-во «Нобелистика», 2008. – С. 80–87.

4. Тютюнник В.М., Сергеев А.Ю. Экспериментальная оценка показателей эффективности сетевого информационного проблемно-ориентированного поиска (на примере нобелистики) // Информатика: проблемы, методология, технологии: материалы 7 междунар. науч.-методолог. конф. – Воронеж: Изд-во ВГУ, 2007. – С. 430–434.

5. Цыганов Н.Л., Циканин М.А. Исследование методов поиска дубликатов веб-документов с учётом запроса пользователя // Интернет-математика-2007: сб. работ участников конкурса. – Екатеринбург: Изд-во Урал. ун-та, 2007. – С. 211–222.

### References

1. Sergeev A.Y., Tyutyunnik V.M. *Internat. Jour. Experiment. Education,* 2012, no.7, pp. 61–66.

2. Sergeev A.Yu., Tyutyunnik V.M. Fundamental Research, 2013, no. 4(2). In press.

3. Sergeev A.Yu., Tyutyunnik V.M. *Materialy 7 VserossijskojMezhvuzovskojNauchnojConferencii* «Formirovanie sprtsialista v usloviyakh regiona» (Materials of 7 All-Russian. Scientific. Conf. Molding of specialist under the conditions of the region: new approaches). Tambov, 2008, pp. 80–87.

4. Tyutyunnik V.M., Sergeev A.Yu. Materialy 7 Mezhdunarodnoj Nauchno-Metodologicheskoj Konferencii (Materials 7th Intern. Sci.-Methodologist. Conf. The Information Theory: Problems, Methodology, Technology). Voronezh, 2007, pp. 430–434.

5. Tsyganov N.L., Tsikanin M.A. SbornikRabotUchastnikovKonkursa «Internet-matematika-2007» (Coll. the work of participants in the competition «Internet-mathematics-2007»). Ekaterinburg, 2007, pp. 211–222.

**Рецензенты:**

Сысоев В.А., д.т.н., профессор кафедры прикладной информатики Тамбовского филиала Московского государственного университета культуры и искусств, г. Тамбов;

Гусятников В.Н., д.ф.-м.н., профессор, заведующий кафедрой прикладной математики и информатики Саратовского государственного социально-экономического университета, г. Саратов.

Работа поступила в редакцию 21.06.2013.