УДК 025.4.036

# МЕТОДИКА ОЦЕНКИ ЭФФЕКТИВНОСТИ ТЕМАТИКО-ОРИЕНТИРОВАННОГО ИНТЕРНЕТ-ПОИСКА С ПОМОЩЬЮ МИНИМИЗАЦИИ ОБЪЁМА ПОИСКОВОЙ ВЫБОРКИ

**Сергеев А.Ю., Тютюнник В.М.**

*Тамбовский филиал ФГБОУ ВПО «Московский государственный университет культуры и искусств», Тамбов, e-mail: vmt@tmb.ru*

Предложены новые и более объективные показатели эффективности тематико-ориентированного интернет-поиска – коэффициент семантического охвата и минимальный объём поисковой выборки, обеспечивающий тематическую полноту поиска. Разработана и протестирована методика оценки эффективности интернет-поиска, оперирующая семантической составляющей результатов поиска на основе коэффициента семантического потенциала поискового термина. Получен комплекс эмпирических данных по оценке эффективности интернет-поиска с использованием предложенной методики. Экспериментально доказано существенное повышение значения минимального объёма поисковой выборки, обеспечивающего тематическую полноту поиска при увеличении значения коэффициента семантического потенциала поискового термина. Результаты эксперимента подтверждаются статистикой поведения пользователей поисковых машин интернета, теоретически неправильного с точки зрения общепринятого подхода. Доказано, что интернет-поиск с использованием тематически сложных поисковых терминов является неэффективным. В результате применения предложенной методики удалось повысить эффективность тематико-ориентированного интернет-поиска более чем в три раза.

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

# METHOD OF ESTIMATION OF SUBJECTS-FOCUSED INTERNET RETRIEVAL EFFICIENCY BY MINIMIZATION OF THE SEARCH SAMPLE VOLUME

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

*Tambov Branch Moscow State University of Culture and Arts, Tambov, e-mail: vmt@tmb.ru*

The new and more objective performance indicators topical-oriented Internet search – semantic coverage ratio and minimum sampling provides a thematic search engine complete search. Developed and tested methods of evaluating the effectiveness of an Internet search, the semantic component of search results with a focus on the basis of semantic potential search term. Received a set of empirical data to assess the effectiveness of an Internet search using the proposed method. Experimentally proved significantly raised the minimum volume of sample search engine that provides a thematic search, when you increase the fullness factor value semantic potential search term. The results of the experiment are confirmed by the statistics of the behavior of users of Internet search engines, it is theoretically incorrect in terms of the conventional approach. Proved that an Internet search using thematically complex search terms is ineffective. As a result of the application of the proposed method has improved the effectiveness of topical-oriented Internet searches more than three times.

**Keywords: information retrieval, efficiency, optimization, precision, recall, semantic potential, factor of semantic coverage, minimum volume of a search sample**

В статьях [2–4] 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. Однако данный показатель оказался сложным в восприятии, а также не универсальным для расчёта эффективности поиска по различным (по коэффициенту семантического потенциала) поисковым терминам.

Очевидный метод работы с тематической составляющей документов основан на использовании семантической (терминологической) модели предметной области. Предлагаемый метод оценки эффективности интернет-поиска в отличие от традиционных не является автономным и требует применения тезауруса или, например, УДК, в качестве семантической модели предметной области. Чем выше полнота представления терминологии, тем более объективным будет оценка предлагаемых показателей.

Тематика поиска раскрывается следующим образом. Запрос приводится к наиболее близкому дескриптору (термину) семантической модели $T_n$, оценивается релевантность документа каждому из нижестоящих дескрипторов $T[k]_{n+1}$, $k$ – их количество. Свойство поискового термина, приведённого к дескриптору тезауруса $T_n$, имеет определённое количество нижестоящих дескрипторов (связанных с исходным связью «ВЫШЕ») $T[k]_{n+1}$ и отражает его лексический и семантический потенциал. Определим *коэффициент семантического потенциала поискового термина k* как количество тематических аспектов наиболее близкого дескриптора семантической модели предметной области, отражённых нижестоящими дескрипторами.

Таким образом, *коэффициент тематического охвата TCF* задаётся нижестоящими дескрипторами по отношению к тематике поискового запроса:

$$TCF_S = \frac{\sum_{1}^{k} T\,[k_{n+1}]}{k \cdot (s - k + 1)}, \qquad (1)$$

$$s \geq \sum_{1}^{k} T\,[k_{n+1}],$$

где $\sum_{1}^{k} T\,[k_{n+1}]$ – количество присутствующих в выборке новых поисковых терминов, отражающих семантический потенциал вышестоящего термина. Параметр $s = [k; \infty]$ задаётся вручную и характеризует объём обрабатываемой поисковой выборки. В случае максимальной эффективности поиска объём поисковой выборки равен количеству нижестоящих дескрипторов.

*Глубина тематического охвата (TCL)* рассчитывается как минимальный объём поисковой выборки, в котором раскрыт семантический потенциал поискового термина $V_{\min}$, т.е. представлена информация по комплексу $T[k]_{n+1}$

$$TCL = \frac{k}{V_{\min}} \text{ при } \sum_{1}^{k} T[k]_{n+1} = k. \qquad (2)$$

Сложность задачи нахождения комплекса $T[k]_{n+1}$ при равном отношении $\frac{k}{V_{\min}}$ увеличивается с увеличением $k$. В таком случае необходимо ввести в формулу (2) коэффициент усложнения задачи.

Для того чтобы определить зависимость *TCL* от значения показателя семантического потенциала поискового термина, а также получить данные о распределении пертинентных документов в поисковой выборке и ряд других сведений, нами проведено объёмное экспериментальное исследование.

### Описание эксперимента

В качестве терминологической модели использовалась УДК. В качестве поискового термина случайным методом выбирался дескриптор $T_n$ с необходимым значением показателя семантического потенциала $k = [1...9]$. Для каждого уровня проводилось семь операций поиска по различным поисковым запросам с помощью ПМ Яндекс для обеспечения надёжности эмпирических данных. Прагматически реальным ограничением объёма исследуемой поисковой выборки принято сто первых найденных документов. Для каждого дескриптора $T[k]_{n+1}$ в поисковой выборке методом экспертной оценки выявлялся первый пертинентный документ и фиксировался его порядковый номер. В случае отсутствия пертинентного документа в поисковой выборке в качестве порядкового номера принималось прагматически реальное значение объёма обрабатываемой поисковой выборки. Статистическая достоверность данных обеспечивалась семью точками (значениями $s$) для каждого количества подуровней. Эксперимент проводился итерационным методом, проведено более трёхсот операций поиска, проанализировано более десяти тысяч найденных документов на предмет пертинентности (отражения одного или нескольких аспектов поискового термина). Последовательность операций изображена на рис. 1.

В результате получены данные о вариации объёма поисковой выборки для комплекса дескрипторов с показателем семантического потенциала $k = [1...9]$. Фрагмент полученных экспериментальным путём данных отражён в таблице.

### Анализ экспериментальных данных

Распределение комплекса пертинентных документов (рис. 2), содержащих новую информацию по теме поиска в поисковой выборке, соответствует данным глобальной поисковой статистики [1, 5].

Эмпирические данные объясняют поведение большинства пользователей ПМ, ограничивающихся вопреки рекомендациям просмотром первой страницы поисковой выборки. В то же время статистические данные, согласующиеся с результатами эксперимента, подтверждают правильное направление в определении методики оценки эффективности поиска. Противоположная ситуация складывается при использовании стандартной методики оценки, основанной, прежде всего, на показателе точности поиска. Динамика точности поиска, полученная в результате исследования, проведённого и описанного в [3], представлена на рис. 3. Постоянная в процессе просмотра поисковой выборки эффективность поиска, основанная на показателе точности, противоречит статистическим данным.

Вероятность присутствия пертинентного документа, содержащего новую информацию (информацию по тематическому аспекту поискового термина, не освещённому в просмотренной выборке), после первых пятнадцати позиций снижается более чем в 5 раз. Таким образом, *несмотря на высокую постоянную точность поиска [3], большинство пользователей ограничивается просмотром лишь первых позиций выборки вследствие снижения вероятности получить новую информацию по теме поиска.*

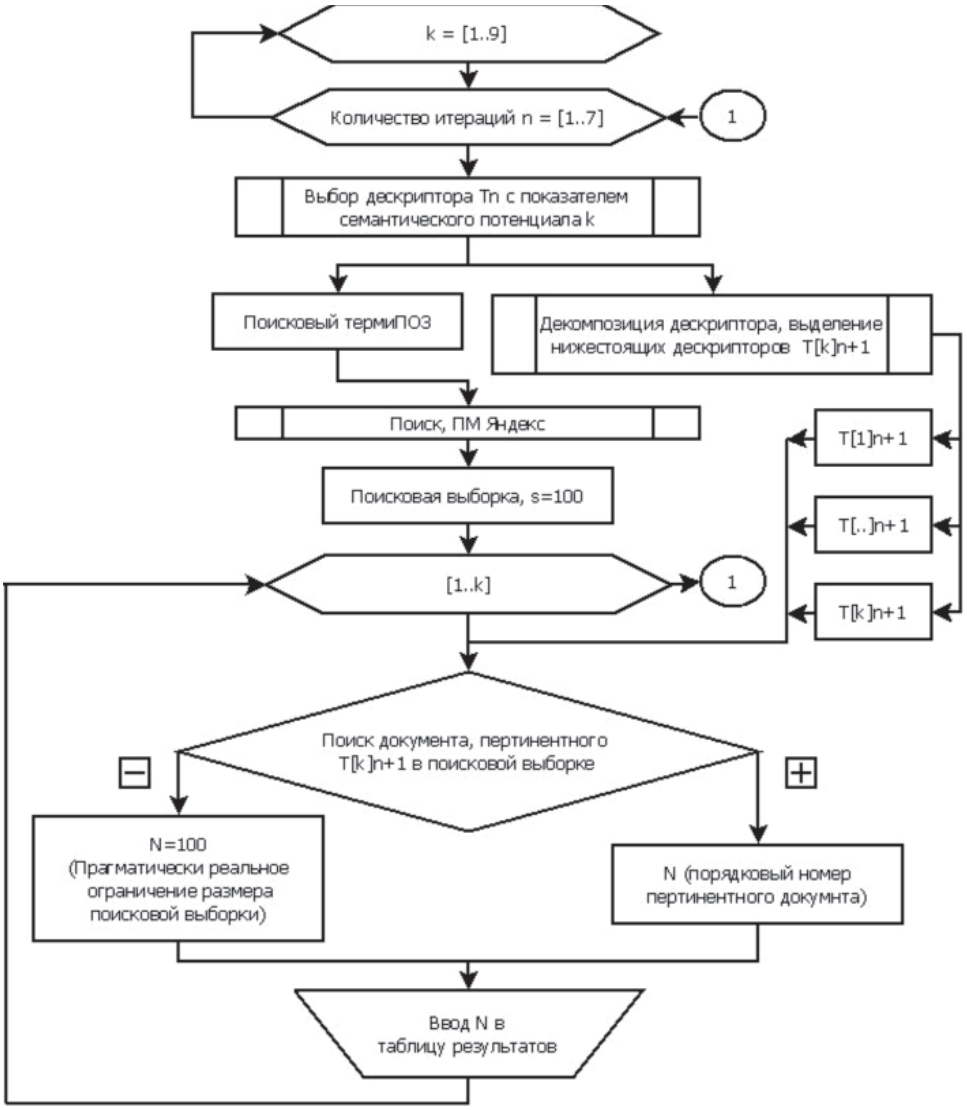


*Рис. 1. Алгоритм получения эмпирических данных*

Фрагмент представления экспериментальных данных

| ***Поисковый термин*** $T[k]_{n+1}$ | | | | Объем поисковой выборки | Порядковый номер пертинентного документа |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 |
| **5 подуровней** | | | | | |
| **Абстрактные типы данных** | | | **s** | **86** | |
| Списки | | | | | 1 |
| Стеки | | | | | 1 |
| Очереди | | | | | 1 |
| Деки, двухсторонние очереди | | | | | 86 |
| Таблицы | | | | | 2 |
| **Синтаксис программ** | | | **s** | **34** | |
| Конкретный синтаксис | | | | | 22 |
| Абстрактный синтаксис | | | | | 32 |
| Регулярное выражение | | | | | 34 |

**Продолжение таблицы**

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Бесконтекстная (контекстно-свободная) грамматика (Смотри также:) | | | | | 9 |
| Атрибутивная грамматика (Смотри также:) | | | | | 24 |
| **Искусственный интеллект** | | | **s** | **38** | |
| Модели когнитивных процессов | | | | | 20 |
| Представление знаний | | | | | 2 |
| Рассуждение | | | | | 2 |
| Обучение | | | | | 18 |
| **Регистры процессоров** | | | **s** | **9** | |
| Регистры общего назначения | | | | | 1 |
| Аккумуляторы. Накапливающие регистры | | | | | 9 |
| Сдвиговые регистры | | | | | 1 |
| Регистры чисел с плавающей запятой | | | | | 5 |
| Стековые регистры | | | | | 2 |
| **Арифметические устройства** | | | **s** | **69** | |
| Сумматоры. Полусумматоры | | | | | 7 |
| Схемы образования дополнения | | | | | 7 |
| Множительные устройства. Делительные устройства | | | | | 58 |
| Векторные арифметические устройства | | | | | 69 |
| Сдвигающие устройства. Сравнивающие устройства | | | | | 67 |
| **Функционирование компьютерной памяти** | | | **s** | **6** | |
| Чтение | | | | | 2 |
| Запись | | | | | 2 |
| Доступ | | | | | 6 |
| Адресация | | | | | 4 |
| Поблочная передача | | | | | 52 |
| Суперкомпьютеры. Супер-ЭВМ | | | **s** | **19** | 2 |
| Миникомпьютеры. Мини-ЭВМ | | | | | 2 |
| Рабочие станции | | | | | 19 |
| Персональные компьютеры. ПЭВМ. Микрокомпьютеры. Микро-ЭВМ | | | | | 19 |
| | **$s_{ср}$ =** | **37,29** | | | - |
| **Стандартное отклонение =** | | **20,31** | | | - |
| **Стандартное отклонение ($s$) =** | | **30,29** | | | |
| **$s > 100, \% =$** | | **0,00** | | | |
| **6 подуровней** | | | | | |
| **Компьютерная графика** | | | **s** | **74** | |
| Элементы и объекты компьютерной графики | | | | | 2 |
| Стереоскопическая визуализация | | | | | 74 |
| Методы ввода графики (Смотри также:) | | | | | 20 |
| Методы компьютерной графики | | | | | 1 |
| Анимация. Мультипликация | | | | | 5 |
| **Структурированные данные. Структуры данных** | | | **s** | **33** | |
| Массивы. Матрицы | | | | | 33 |
| Записи | | | | | 4 |
| Множества | | | | | 11 |
| Динамические структуры данных | | | | | 21 |
| Абстрактные типы данных | | | | | 11 |
| Другие типы данных | | | | | 23 |
| **Сети ЭВМ. Вычислительные сети** | | | **s** | **26** | |
| Архитектура сетей (Сетевые протоколы – Смотри:) | | | | | 3 |
| Виды сетей в зависимости от охватываемой территории | | | | | 3 |

**Окончание таблицы**

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Применение компьютерных сетей в целом. Применение интернета | | | | | 3 |
| Диалоговые вычислительные системы для специальных целей | | | | | 26 |
| **Световое перо** | | | **s** | **100** | **3** |
| Сенсорные экраны | | | | | 3 |
| Мышь | | | | | 4 |
| Шары трассировки | | | | | 100 |
| Рычажные указатели. Джойстики | | | | | 8 |
| **Абстракция** | | | **s** | **100** | **12** |
| Разбиение на модули. Модуляризация | | | | | 35 |
| Упрятывание информации | | | | | 100 |
| Программирование отдельных компонентов системы | | | | | 100 |
| Языки проектирования программ. Псевдокод. Символический код | | | | | 32 |
| Спецификация проекта программного обеспечения | | | | | 24 |
| **Сетевая аппаратура** | | | **s** | **100** | |
| Сетевые адаптеры. Сетевые платы | | | | | 1 |
| Коммутаторы данных | | | | | 1 |
| Маршрутизаторы | | | | | 1 |
| Устройства сетевой связи. Мосты, шлюзы, реле | | | | | 1 |
| **Представление знаний** | | | **s** | **100** | |
| Сети знаний. Семантические сети | | | | | 2 |
| Фреймовые системы. Фреймы. Схемы. Сценарии | | | | | 2 |
| Множественные миры | | | | | 4 |
| Порождающие системы. Системы правил вывода | | | | | 2 |
| Модель чёрной доски | | | | | 100 |
| **$s_{ср}$ =** | | **76,14** | | | - |
| **Стандартное отклонение =** | | **32,38** | | | - |
| **Стандартное отклонение ($s$) =** | | **33,31** | | | |
| **$s$ > 100, % =** | | **57,00** | | | |

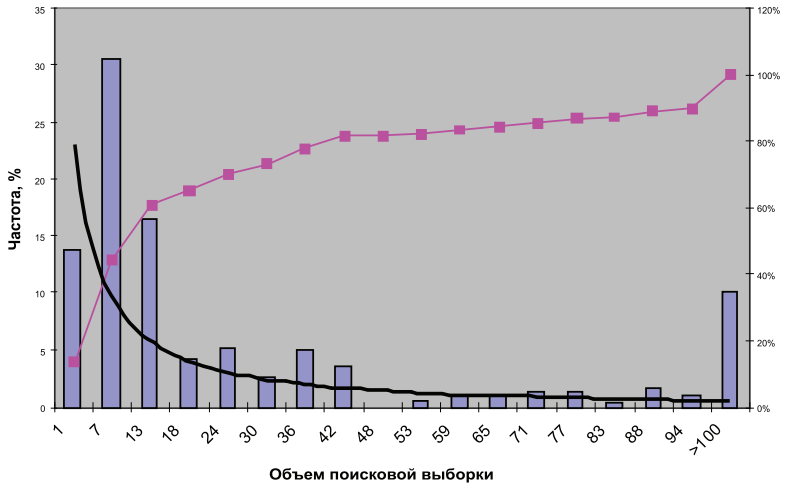


*Рис. 2. Зависимость частоты распределения пертинентных страниц, содержащих новую информацию, от объёма поисковой выборки*

Эксперимент показал, что эффективность поиска зависит от тематической сложности поискового термина, оцениваемой с помощью коэффициента семантического потенциала. *Повышение тематической сложности поискового термина ведёт к снижению эффективности поиска при постоянном значении точности.* Другими словами, в процессе движения вглубь поисковой выборки при неизменной точности поиска снижается доля новой информации. Именно с этим фактом

снижения эффективности поиска связано нежелание пользователей ПМ просматривать всю поисковую выборку, даже если искомая информация не найдена. В таком случае пользователь прибегает к модификации запроса, что оказывается более эффективным, чем продолжать просматривать результаты поиска.

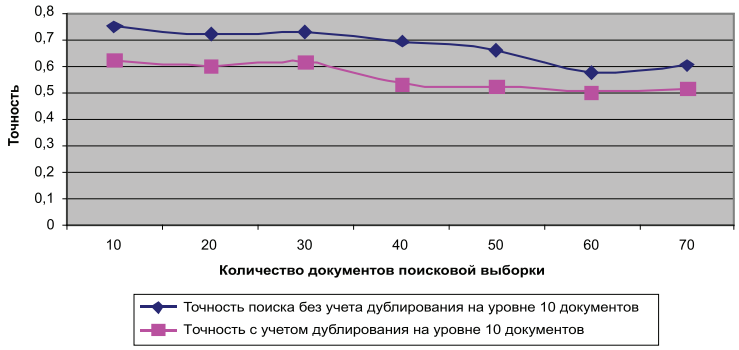


*Рис. 3. Динамика показателя точности поиска от количества документов в поисковой выборке*

В ходе анализа данных эксперимента выявлена эмпирическая зависимость между значением объёма поисковой выборки и значением коэффициента семантического охвата поискового термина (рис. 4).

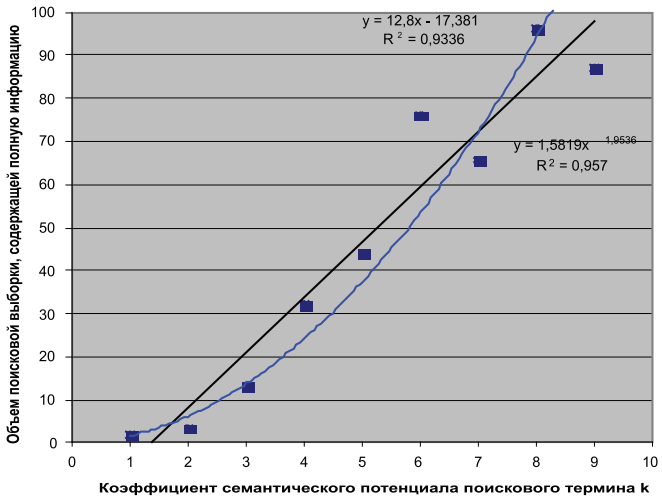


*Рис. 4. Зависимость глубины поискового охвата от коэффициента семантического охвата поискового термина*

Найденная зависимость показывает быстрое снижение тематической полноты поиска и увеличение объёма поисковой выборки с увеличением значения коэффициента семантического потенциала. Эмпирические данные аппроксимируется как степенной (3), так и линейной (4) функциями со сходной достоверностью аппроксимации:

$$V_{\min} = 1{,}6k^2; \qquad (3)$$

$$V_{\min} = 12{,}8k - 17{,}4. \qquad (4)$$

На основе полученных эмпирических закономерностей включим в формулу расчёта *TCL* коэффициент усложнения задачи нахождения комплекса $T[k]_{n+1}$ при равном отношении с увеличением $k$. В этом случае формула (2) для стандартной методики поиска примет вид:

$$TCL = \frac{12{,}8k - 17{,}4}{V_{\min}},\ k > 1;$$

$$TCL = 1{,}6\frac{k^2}{V_{\min}}.$$

На данном этапе исследования принято решение о неудобстве восприятия и расчёта показателя тематической глубины поиска, сравнения эффективности поиска при использовании терминов с различным значением коэффициента семантического потенциала. Логически правильнее оценивать *минимальный объём поисковой выборки, в котором раскрыт семантический потенциал поискового термина* $V_{\min}$.

Как показал эксперимент, тематическая полнота не обеспечивается объёмом поисковой выборки в 100 документах при значении коэффициента тематического потенциала поискового термина $k > 5$ (рис. 4). Здесь мы видим линейную зависимость, которая свидетельствует о повышении доли поисковых сессий, при которых не полностью освещена тематика поиска, при повышении коэффициента тематического потенциала поискового термина. Учитывая разброс пертинентных документов, содержащих новую информацию по тематике поиска, и высокое процентное соотношение поисковых сессий, при которых тематика поиска не освещена полностью (рис. 5), экспериментальное исследование показало, что *интернет-поиск с использованием тематически сложных поисковых терминов является неэффективным*.

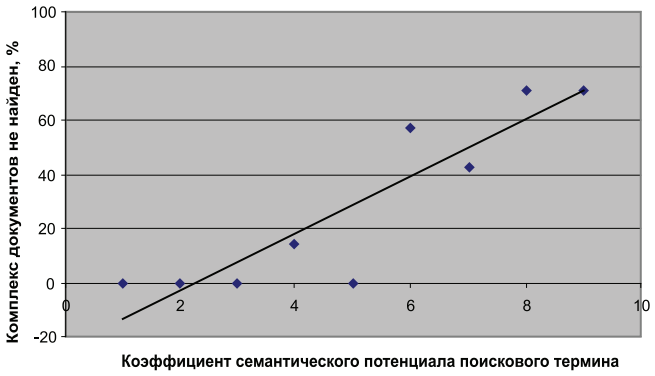


*Рис. 5. Зависимость количества поисковых сессий, при которых тематика поиска не освещена полностью, от семантического потенциала поискового термина*

## Выводы

1. Предложен новый показатель эффективности интернет-поиска – минимальный объём поисковой выборки, раскрывающий семантический потенциал поискового термина (обеспечивающий тематическую полноту поиска).

2. Предложена новая методика оценки эффективности интернет-поиска, оперирующая семантической составляющей результатов поиска на основе коэффициента семантического потенциала поискового термина.

3. Получен комплекс эмпирических данных по оценке эффективности интернет-поиска с использованием предложенной методики.

4. Экспериментально доказано существенное повышение значения минимального объёма поисковой выборки, обеспечивающего тематическую полноту поиска при увеличении значения коэффициента семантического потенциала поискового термина. Результаты эксперимента подтверждаются статистикой поведения пользователей ПМ, теоретически «неправильного» с точки зрения общепринятого подхода.

5. Показано, что тематико-ориентированный интернет-поиск с использованием поисковых терминов с показателем семантического потенциала $k > 3$ является низкоэффективным.

### Список литературы

1. Исследование модели поведения пользователей при работе с поисковыми системами. – URL: http://optimization.ru/articles/traffic2007 (дата обращения: 12.04.12).

2. Сергеев А.Ю., Тютюнник В.М. Эффективность тематико-ориентированного Интернет-поиска // Международный журнал экспериментального образования. – 2012. – № 7. – С. 61–66.

3. Сергеев А.Ю., Тютюнник В.М. 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.

4. Сергеев А.Ю., В.М.Тютюнник. 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. / под ред. проф. В.М. Тютюнника, В.А. Сысоева. – Тамбов; М.; СПб.; Баку; Вена: Изд-во «Нобелистика», 2008. – С. 80–87.

5. Alemayehu N. Analysis of Performance Variation Using Query Expansion // Journal of the American Society for Information Science and Technology. – 2003. – № 54(5). – P. 379–391.

### References

1. Study on models of user behavior when working with search engines. URL: http://optimization.ru/articles/traffic2007 (date: 04/12/12).

2. Sergeev A.Ju, Tyutyunnik V.M. International Journal of experimental education, 2012, no.7, pp. 61–66.

3. Sergeev A.Ju., Tyutyunnik V.M. Information science: problems, methodology, technology: materials 7 intern. researcher-expert conf. Voronezh, 2007, pp. 430–434.

4. Sergeev A.Ju., Tyutyunnik V.M. A specialist formation in the region condition: new approaches: materials of 7th Vseros. mezhvuz. nauch. konf. Tambov, 2008, pp. 80–87.

5. Alemayehu N. Analysis of Performance Variation Using Query Expansion // Journal of the American Society for Information Science and Technology. 2003. no. 54(5). pp. 379–391.

**Рецензенты:**

Гусятников В.Н., д.ф.-м.н., профессор, заведующий кафедрой прикладной информатики и математики, ФГБОУ ВПО «Саратовский государственный социально-экономический университет», г. Саратов;

Громов Ю.Ю., д.т.н., профессор, декан факультета информационных технологий, ФГБОУ ВПО «Тамбовский государственный технический университет», г. Тамбов.

Работа поступила в редакцию 04.02.2013.