УДК 004.051

С.Е. Савотченко, В.А. Стукалов, Е.А. Проскурина

ДИНАМИКА СЕМАНТИЧЕСКОЙ МЕРЫ РЕЗУЛЬТАТОВ ПОИСКОВЫХ ЗАПРОСОВ

Савотченко Сергей Евгеньевич, *доктор физико-математических наук, доцент, окончил физический факультет Харьковского государственного университета, в настоящее время профессор кафедры «Информационные технологии» Белгородского института развития образования. Имеет статьи в области математического моделирования, информационных технологий и автоматизированных информационных систем. [e-mail: savotchenko@hotbox.ru].*

Стукалов Вадим Андреевич, *окончил Белгородский государственный университет, аспирант кафедры «Библиотековедение, библиографоведение и книговедение» Белгородского государственного института искусств и культуры. Имеет статьи в области автоматизированных библиотечных информационных систем. [e-mail: svadglo@gmail.com].*

Проскурина Елена Александровна, *окончила Белгородский государственный институт искусств и культуры, аспирантка кафедры «Библиотековедение, библиографоведение и книговедение» Белгородского государственного института искусств и культуры. Имеет статьи в области автоматизированных библиотечных информационных систем. [e-mail: pea@bgiik.ru].*

Аннотация

В работе приведены результаты исследования изменения с течением времени характеристик качества информационного поиска на примере пяти поисковых систем глобальной сети. Для выявления возможности выдавать автоматизированными поисковыми системами пертинентные документы при простой форме поиска были рассчитаны реализации семантической меры и парциальной семантической меры результатов поисковых запросов понятия.

Для анализа устойчивости с течением времени результатов информационного поиска в работе использовались количественные показатели, характеризующие выполнение последовательности условно нормализованных запросов. Определена методика проведения исследований, позволяющих зафиксировать динамику семантической меры результатов запросов к автоматизированным поисковым информационным системам.

В статье приводится динамика показателей полноты семантических связей и семантической меры поисковой выдачи систем информационного поиска yandex.ru, rambler.ru, nigma.ru, qip.ru, mail.ru. На основании реализаций траекторий случайных зависимостей семантической меры и парциальной меры рассчитаны средние по времени наблюдений значения. Анализ вариативности значений парциальной меры как функции веса показал, что среди обследованных информационно-поисковых систем mail.ru обладает в определенной степени возможностью проведения пертинентного поиска.

Ключевые слова: дескриптор, информационный поиск, семантические связи, парадигматические отношения, информационно-поисковые системы.

DYNAMICS OF SEMANTIC MEASURE OF SEARCH QUERY RESULTS

Sergey Evgenyevich Savotchenko, *Doctor of Physics and Mathematics, Associate Professor; graduated from the Faculty of Physics at Kharkiv State University; Professor of the Department of Computer Science in Belgorod Institute of Education Development, an author of articles in the field of mathematical modeling, information technology, and automated information systems. e-mail: savotchenko@hotbox.ru.*

Vadim Andreevich Stukalov, *Postgraduate Student of the Department of Library Science, Bibliography and Bibliology at the Belgorod State Institute of Art and Culture; graduated from Belgorod State University; an author of articles in the field of the integrated library systems. e-mail: svadglo@gmail.com.*

Elena Aleksandrovna Proskurina, *Post-graduate Student of the Department of Library Science, Bibliography and Bibliology at the Belgorod State Institute of Art and Culture; graduated from the Belgorod State Institute of Art and Culture; an author of articles in the field of the integrated library systems. e-mail: pea@bgiik.ru.*

Abstract

The article presents the research results of changing the quality characteristics of information search in time illustrated by the example of five global search systems. The implementation of semantic measure and partial semantic measure of definition search query results was estimated to display that the automated search systems are in a position to output the pertinent documents at simple select query.

To analyze the stability over time the results of information retrieval, in the use of quantitative indicators for implementation of the sequence of normalized conditional queries. Defined methodology for conducting trials to capture the dynamics of semantic query results to measure the automated retrieval information systems.

We used the quantitative characteristics specifying the conditional standard query sequence implementation to analyze the stability of retrieval results in time. It was defined the methodology for conducting trials to capture the dynamics of semantic query results to measure the automated retrieval information systems.

The article presents dynamics of completeness semantic relations and semantic measures of the SERPs information retrieval systems Yandex, Rambler, Nigma, Qip, and Mail.ru. Based on the realizations of random trajectories semantic dependency measures and partial measures, mean values are calculated for the observation period. Analysis of variance values of the partial measures as a function of weight showed that among the surveyed search engines, Mail.ru has to some extent the possibility of pertinence search.

Key words: descriptor, information search, semantic relations, paradigmatic relations, information retrieval systems.

Введение

В последнее время начало развиваться направление, связанное с улучшением не столько показателей релевантности [1, 2], а скорее пертинентности результатов поисковых запросов в информационно-поисковых системах (ИПС). Реализация такого направления ведется различными способами, одними из которых являются автоматизация построения тезаурусов, использование семантических сетей и онтологий. Исследования по выявлению методик повышения качества информационного поиска с помощью онтологии ведутся по двум направлениям – применение семантических технологий в сети Интернет и электронных библиотеках. В электронных библиотеках уже активно используются методы и средства онтологического моделирования пространств библиотечных знаний. В рамках этого направления уже много лет ведутся исследования в области анализа семантики связей между данными, по которым осуществляется поиск. Системным обобщением этих результатов стало появление комплекса онтологий SPAR, а также появление семантического раздела в модели научных данных CERIF. В этом направлении ведется серьезная работа и в рамках консорциума W3C, где в проекте SKOS (Simple Knowledge Organization System) предлагается модель связывания научных данных, адаптированная для компьютерной обработки. В частности, SKOS включает контролируемые структурные словари семантических значений для связывания научных данных [3]. Резко возросло практическое применение онтологий в сети Интернет. Например, онтологии используются в Google для классификации веб-сайтов. Компания Amazon разработала онтологию товаров и услуг с их характеристиками. Другой пример – онтология UNSPSC (United Nations Standard Products and Services Code – система ООН стандартных кодов для товаров и услуг) [4]. Кроме того, проводятся разработки моделей информационного поиска с учетом онтологий [5, 6], семантических связей [7].

Количественные показатели

Для анализа устойчивости с течением времени результатов информационного поиска целесообразно использовать количественные показатели, характеризующие выполнение последовательности условно нормализованных запросов, все члены которой связаны четкими парадигматическими отношениями: $Q = \{\text{д}, \text{с}, \text{вр}, \text{вц}, \text{нч}, \text{нв}, \text{а}\}$, где обозначаются: (д) – заглавный дескриптор – ведущая лексическая единица (ЛЕ); (с) – синоним к ведущей ЛЕ; (вр) – вышестоящее родовое понятие к ведущей ЛЕ; (вц) – вышестоящее целое понятие к ведущей ЛЕ; (нч) – нижестоящее частное понятие к ведущей ЛЕ; (нв) – нижестоящее видовое понятие к ведущей ЛЕ; (а) – ассоциация с ведущей ЛЕ.

Используются величины [8–10]:

1) Объем i-го уровня запроса – количество результатов поиска, то есть документов, выдаваемых ИПС S_l на i-ю ЛЕ последовательности запросов Q в: $A_i = A_i(Q, S_l)$.

2) Индекс i-го и j-го уровней – отношение объема i-го уровня к объему j-го уровня:

$$J_{ij}(Q, S_l) = A_i(Q, S_l) / A_j(Q, S_l). \quad (1)$$

Оптимальным набором показателей является группа индексов $J_k = \{J_{10}, J_{20}, J_{30}, J_{40}, J_{50}, J_{60}, J_{23}, J_{45}, J_{16}\}$.

Поскольку в различные моменты времени результаты информационного поиска по одному и тому же запросу могут отличаться, а результат выполнения запроса заранее предсказать нельзя, то величины $A_i = A_i(t)$ и $J_{ij} = J_{ij}(t)$ и представляют собой случайные процессы. В результате проведения одного и того же запроса в различные моменты времени можно получить реализацию соответствующего случайного процесса.

Как было показано на примере ИПС Google в [10], результаты информационного поиска испытывают колеба-

ния, а величины каждого показателя группируются около определенных средних значений:

$$\overline{J}_{ij} = \frac{1}{n}\sum_{k=1}^{n} J_{ij}\left(t_k\right), \quad (2)$$

где n – количество моментов наблюдений t_k.

Перед вычислением средних значений рекомендуется исключить явно «выскакивающие данные» (провести сглаживание рядов динамики). Для оценки статистической погрешности результатов сначала необходимо вычислить исправленные дисперсии:

$$s_{ij}^2 = \sum_{k=1}^{n}\left(J_{ij}\left(t_k\right) - \overline{J}_{ij}\right)^2 \Big/ \left(n-1\right). \quad (3)$$

Затем следует вычислить абсолютную погрешность для каждого среднего:

$$\delta_{ij} = t_\gamma\left(n\right) s_{ij} \Big/ \sqrt{n}, \quad (4)$$

где $t_\gamma(n)$ – значение, определяемое из специальных статистических таблиц для заданного количества наблюдений n и доверительной вероятности γ [11]. Относительные статистические погрешности вычисляются по формуле:

$$\varepsilon_{ij} = \frac{\delta_{ij}}{\overline{J}_{ij}} \cdot 100\%, \quad (5)$$

В работе [12] введено понятие семантической меры результатов поисковых запросов – взвешенное среднее гармоническое значение показателей полноты семантических связей (1):

$$F_s\left(Q_m, S_l\right) = 1 \Big/ \sum_k \nu_k \Big/ J_k\left(Q_m, S_l\right), \quad (6)$$

где ν_k – веса, такие что $\sum_{k=1}^{n} \nu_k = 1$, n – количество усредняемых показателей, суммирование производится по выделенной группе пар чисел, например, $k = \{1\ 0, 2\ 0, 3\ 0, 4\ 0, 5\ 0, 6\ 0, 2\ 3, 4\ 5, 1\ 6\}$, тогда $n = 9$.

В [12] был также введен специальный вариант меры (2), названный парциальной или частичной:

$$F_{16}\left(Q_m, S_l\right) = \frac{1}{\dfrac{\nu}{J_{10}\left(Q_m, S_l\right)} + \dfrac{1-\nu}{J_{60}\left(Q_m, S_l\right)}}, \quad (7)$$

где вес $\nu \in [0;\ 1]$, представляющий собой аналог F-меры (меры Ван Ризбергена) [1, 2].

Поскольку мера F_{16} определяется через показатели (1), то так же, как и они, она представляет собой случайный процесс $F_{16}(t)$. Его среднее по времени значение:

$$\overline{F}_{16} = \frac{1}{n}\sum_{k=1}^{n} F_{16}\left(t_k\right). \quad (8)$$

Для оценки степени изменения значения F_{16} как функции веса ν на интервале $[0;\ 1]$ можно использовать относительный размах варьирования:

$$\delta\overline{F}_{16} = \frac{R}{\overline{F}_{16}} \cdot 100\%, \quad (9)$$

где $R = F_{16\max} - F_{16\min}$ – размах варьирования.

Если величина (9) не превосходит 5%, то можно считать, что изменение соответствующего показателя в диапазоне наблюдений практически не происходит.

Методика проведения исследования

Для проведения исследований выбран ряд русскоязычных ИПС: $S_1 = \{\text{yandex.ru}\}$, $S_2 = \{\text{rambler.ru}\}$, $S_3 = \{\text{nigma.ru}\}$, $S_4 = \{\text{qip.ru}\}$, $S_5 = \{\text{mail.ru}\}$. Используемая последовательность запросов, сформированная по тезаурусу (ГОСТ 7.25-2001), имеет вид: $Q_{(\text{д})} = \{\text{обучение}\}$, $Q_{(\text{с})} = \{\text{воспитание}\}$, $Q_{(\text{вр})} = \{\text{педагогический процесс}\}$, $Q_{(\text{вц})} = \{\text{образование}\}$, $Q_{(\text{нч})} = \{\text{заочное обучение}\}$, $Q_{(\text{нв})} = \{\text{лекционное занятие}\}$, $Q_{(\text{а})} = \{\text{ученик}\}$.

Методика проведения исследований следующая. В строке поиска ИПС вводится первая ЛЕ последовательности $Q_{(\text{д})}$. Количество выданных по этому запросу документов есть величина A_1. Затем в этой же ИПС вводится второй член последовательности $Q_{(\text{с})}$. Количество выданных по этому запросу документов есть величина A_2. И так далее для всех членов последовательностей всех запросов последовательности Q, в результате чего получается необходимый набор объемов A_i. Затем с помощью этих величин вычисляются индексы по формуле (1). Далее вся процедура повторяется через определенные интервалы времени.

В результате получается целый набор данных показателей для различных моментов времени, которые представляют собой реализации соответствующих случайных процессов. Полученные данные используются для вычисления описанных выше количественных показателей.

Результаты

Исследование проводилось в период с 15.10.2012 по 15.02.2013, запросы по ИПС проводились два раза в неделю. На рисунке 1 для иллюстрации представлены графики полученных реализаций процессов $J_{10}(t)$ и $J_{60}(t)$ для рассматриваемых ИПС. Аналогичным образом ведут себя и остальные показатели выделенного набора. Уставлено, что все они испытывают случайные осцилляции. Их значения группируются около средних, сводные диаграммы для которых приведены на рисунке 2 в виде распределения по ИПС.

В ходе проведения исследований было произведено $n = 34$ наблюдений в указанный период времени, тогда для $\gamma = 0{,}95$ [11]: $t_{0,95}(34) \approx 2{,}034$. Были рассчитаны дисперсии (3), абсолютные (4) и относительные погрешности (5), значения которых представлены в таблице 1. Видно, что относительные статистические погрешности всех показателей для всех исследуемых ИПС не превосходят 5%, что свидетельствует об удовлетворительной точности проведенных наблюдений.

Для полученных данных были рассчитаны реализации семантической меры результатов поисковых запросов (6) как функции моментов наблюдений $F_s = F_s(t)$ для

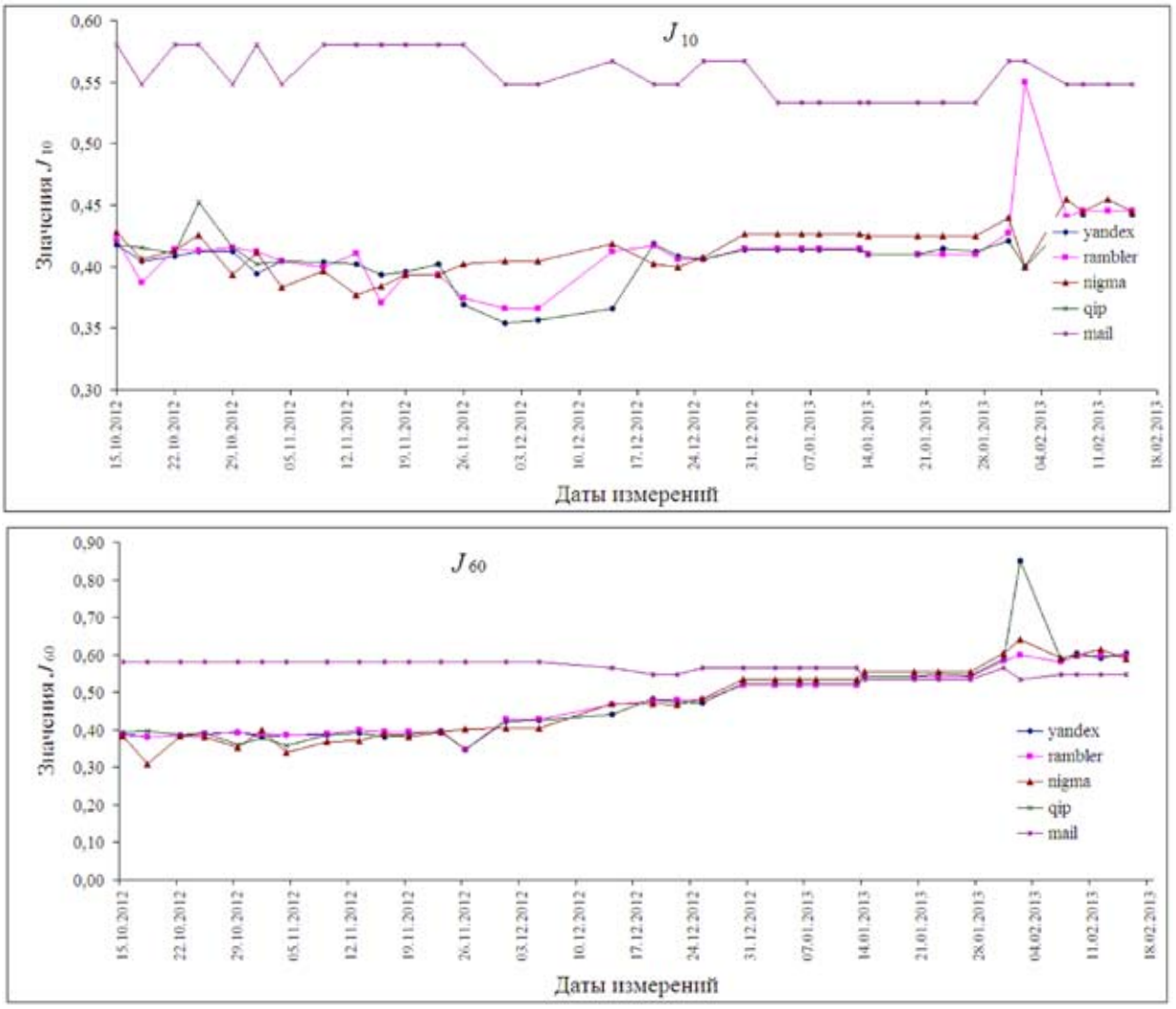

Рис. 1. Графики реализаций процессов $J_{10}(t)$ и $J_{60}(t)$

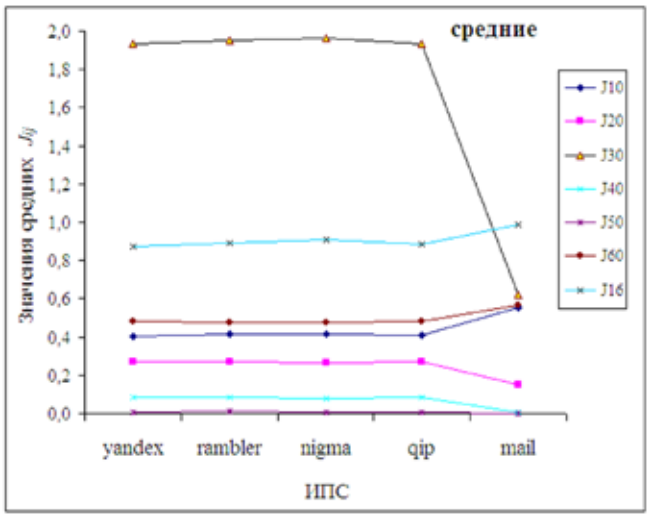

Рис. 2. Распределения средних значений величин J_{ij} по ИПС

Таблица 1

Относительные погрешности ε_{ij} (%)

$\overline{J}_{ij}$	S_1	S_2	S_3	S_4	S_5
$\overline{J}_{10}$	0,572071	0,879804	0,556879	0,611195	0,385214
$\overline{J}_{20}$	3,99089	4,037449	4,103236	3,984824	1,279071
$\overline{J}_{30}$	0,576717	0,65288	0,879871	0,55721	1,825191
$\overline{J}_{40}$	0,670984	0,771252	0,664444	0,87108	0,524845
$\overline{J}_{50}$	3,436848	1,381698	0,697478	3,589726	0,705905
$\overline{J}_{60}$	2,51058	1,994593	2,340399	2,545242	0,362953
$\overline{J}_{23}$	4,13951	4,292134	4,343087	4,109323	1,217297
$\overline{J}_{45}$	2,94128	1,206452	0,864899	3,333453	1,058503
$\overline{J}_{16}$	1,997931	1,646821	2,042747	2,133259	0,336606

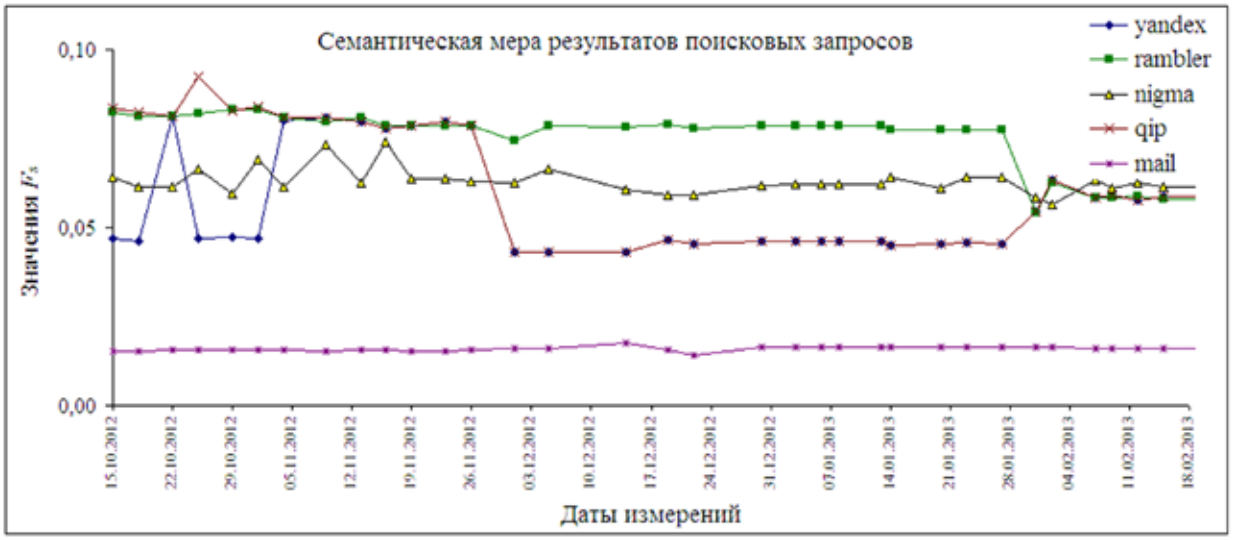

Рис. 3. Динамика семантической меры результатов поисковых запросов

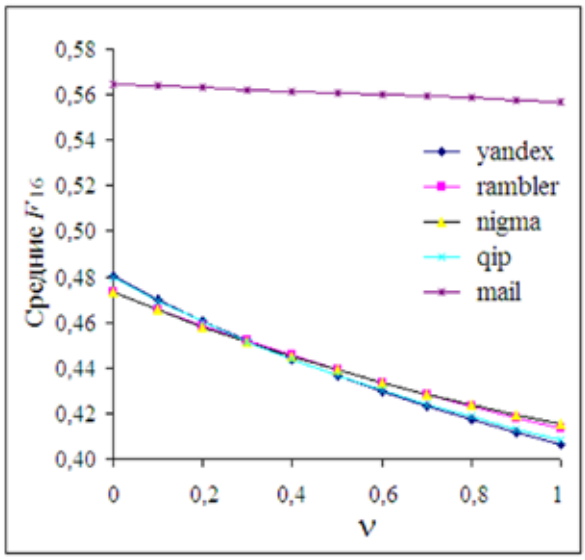

Рис. 4. Зависимости средних значений парциальной меры от веса

Таблица 2

Характеристики вариативности значений парциальной меры

	S_1	S_2	S_3	S_4	S_5
F_{16max}	0,4806621	0,4734809	0,4731216	0,4795985	0,56470588
F_{16min}	0,406143	0,4132517	0,4151614	0,4080829	0,55702087
R	0,0745191	0,0602292	0,0579602	0,0715156	0,00768501
$\overline{F}_{16}$	0,4392127	0,4408846	0,4411073	0,4395688	0,56077814
$\delta\overline{F}_{16}$	16,97%	13,66%	13,14%	16,27%	1,37%

рассматриваемых ИПС. Графики ее реализаций приведены на рисунке 3. Установлено, что наиболее устойчивое поведение демонстрирует реализация $F_s(t)$ для ИПС S_5.

Для выявления способности выдавать пертинентные документы при простой форме поиска были рассчитаны реализации парциальной меры результатов поисковых запросов (7) как функции моментов наблюдений $F_{16} = F_{16}(t)$, а затем найдены их средние значения (8) для рассматриваемых ИПС. На рисунке 4 приведены графики зависимостей средних значений парциальной меры как функции веса $\overline{F}_{16} = \overline{F}_{16}(\nu)$. Видно, что такие зависимости для ИПС $S_1 - S_4$ близки друг к другу, а для S_5 линия выделяется обособленно.

Для приведенных на рисунке 4 зависимостей для всех рассматриваемых ИПС на всем интервале изменения веса $\nu \in [0;\ 1]$ были рассчитаны максимальные и минимальные значения, размахи варьирования R и относительные размахи варьирования (9).

Анализ результатов зависимостей $\overline{F}_{16} = \overline{F}_{16}(\nu)$ показал (см. табл. 2), что только для ИПС S_5 справедливо неравенство $\delta\overline{F}_{16}(S_5) < 5\%$. Следовательно, можно считать, что изменения среднего значения показателя $\overline{F}_{16}(S_5)$ во всем интервале $\nu \in [0;\ 1]$ практически не происходит. Другими словами, величина $\overline{F}_{16}(S_5)$ не зависит от веса. На основании этого можно сделать вывод о том, ИПС S_5 в большей степени способна выдавать пертинентные документы по запросу, чем остальные исследуемые ИПС.

Заключение

Показана динамика показателей полноты семантических связей и семантической меры результатов поисковых запросов на примере ИПС yandex.ru, rambler.ru, nigma.ru, qip.ru, mail.ru. На основании реализаций траекторий случайных зависимостей семантической меры и парциальной меры рассчитаны средние по времени наблюдений значения. Анализ вариативности значений парциальной меры как функции веса показал, что среди обследованных ИПС mail.ru обладает в определенной степени возможностью проведения пертинентного поиска.

СПИСОК ЛИТЕРАТУРЫ

1. Manning Ch. D., Raghavan P., Schütze H. *Vvedeniye v informatsionnyy poisk: per. s angl.* [Introduction to Information Retrieval, Translation from English]. Moscow, Izdatelskiy Dom Williams Publ., 2011. 528 p.

2. Официальные метрики РОМИП'2010. – URL: http://romip.ru/romip2010/20_appendix_a_metrics.pdf.

3. Хорошевский В.Ф. Семантические технологии: ожидания и тренды // Открытые семантические технологии проектирования интеллектуальных систем : матер. II-й Межд. научн.-техн. конф. – Минск : БГУИР, 2012. – С. 143–158.

4. Плесневич Г.С. Формальные онтологии // Открытые семантические технологии проектирования интеллектуальных систем : матер. II-й Межд. научн.-техн. конф. – Минск : БГУИР, 2012. – С. 163–168.

5. Стулов А. Особенности построения информационных хранилищ // Открытые системы. – 2003. – № 4. – URL: http://www.osp.ru/os/2003/04/182942.

6. Селяев А.Г. Взвешивание терминов в процессах индексирования электронных информационных ресурсов // Автоматизация процессов управления. – 2007. – № 2 (10). – С. 93–96.

7. Савотченко С.Е., Жуков П.С. Моделирование информационного поиска в базе данных с учетом семантических связей // Автоматизация процессов управления. – 2013. – № 2 (32). – С. 17–22.

8. Савотченко С.Е., Логинова А.Е. Математический метод сравнительного анализа семантических особенностей информационно-поисковых систем // Теория и практика общественного развития. – 2012. – № 6. – С. 101–104.

9. Савотченко С.Е., Проскурина Е.А. Корреляционный и дисперсионный анализ лингвистических особенностей поиска в интернете // Среднее профессиональное образование. – 2012. – № 12. – С. 38–40.

10. Савотченко С.Е., Проскурина Е.А. Показатели семантических связей информационно-поисковых систем // Научные ведомости «БелГУ». Сер. История. Политология. Информатика. – 2013. – Вып. 25/1, № 1 (144). – С. 145–151.

11. Математическая статистика: учебник для вузов / В.Б. Горяинов [и др.]; под ред. В.С. Зарубина, А.П. Крищенко. – М. : Изд-во МГТУ им. Н.Э. Баумана, 2001. – 424 с.

12. Савотченко С.Е., Стукалов В.А. Семантическая мера результатов поисковых запросов // Автоматизация процессов управления. – 2013. – № 4 (34) – 2013. – С. 57– 60.

REFERENCES

1. Manning Ch. D., Raghavan P., Schütze H. *Vvedeniye v informatsionnyy poisk : per. s angl.* [Introduction to Information Retrieval]. Moscow, ID Williams Publ., 2011. 528 p.

2. Ofitsialnyye metriki ROMIP'2010. [ROMIP'2010 Registered Metrics]. Available at: http://romip.ru/romip2010/20_appendix_a_metrics.pdf.

3. Khoroshevskiy V.F. Semanticheskiye tekhnologii: ozhidaniya i trendy [Semantic Technologies: Expectations and Trends]. *Otkrytyyesemanticheskiyetekhnologiiproyektirovaniya intellektualnykh sistem : mater. II Mezhd. nauchn.-tekhn. Konf* [Open Semantic Technologies for Intelligence Systems: Materials of the II International Scientific Conference], Minsk. BGUIR Publ., 2012, pp. 143–158.

4. Plesnevich G.S. Formalnyye ontologii [Formal Ontologies]. *Otkrytyye semanticheskiye tekhnologii proyektirovaniya intellektualnykh sistem : mater. II Mezhd. nauchn.-tekhn. konf.* [Open Semantic Technologies for Intelligence Systems: Materials of the II International Scientific Conference], Minsk, BGUIR Publ., 2012, pp. 163–168.

5. Stulov A. Osobennosti postroyeniya informatsionnykh khranilishch [Features of Data Warehousing]. *Otkrytyye sistemy* [Open Systems Journal], 2003, no. 4. Available at: http://www.osp.ru/os/2003/04/182942.

6. Selyaev A.G. Vzveshivaniye terminov v protsessakh indeksirovaniya elektronnykh informatsionnykh resursov [Weighting of Terms during Online Information Resources Indexing]. *Avtomatizatsiya protsessov upravleniya* [Automation of Control Processes], 2007, no. 2 (10), pp. 93–96.

7. Savotchenko S.E., Zhukov P.S. Modelirovaniye informatsionnogo poiska v baze dannykh s uchetom semanticheskikh svyazey [Modeling of Information Retrieval in the Database with Semantic Relations]. *Avtomatizatsiya protsessov upravleniya* [Automation of Control Processes], 2013, no. 2 (32), pp. 17–22.

8. Savotchenko S.E., Loginova A.E. Matematicheskiy metod sravnitelnogo analiza semanticheskikh osobennostey informatsionno-poiskovykh sistem [Mathematical Method of the Comparative Analysis of Information-Retrieval Systems' Semantic Features]. *Teoriya i praktika obshchestvennogo razvitiya* [Theory and Practice Social Development], 2012, no. 6, pp. 101–104.

9. Savotchenko S.E., Proskurina E.A. Korrelyatsionnyy i dispersionnyy analiz lingvisticheskikh osobennostey poiska v internete [Correlative and Variance Analysis of Web Search' Linguistic Features]. *Sredneye professionalnoye obrazovaniye* [The Journal of Secondary Vocational Education], 2012, no. 12, pp. 38–40.

10. Savotchenko S.E., Proskurina E.A. Pokazateli semanticheskikh svyazey informatsionno-poiskovykh sistem [Semantic Relations of Information Retrieval Systems]. *Nauchnyye vedomosti BelGU. Ser. Istoriya. Politologiya. Informatika* [Scientific Records of Belgorod State University. History, Politology, Informatics], 2013, vol. 25(1), no. 1 (144), pp. 145–151.

11. Goryainov V.B. and Others. *Matematicheskaya statistika: uchebnik dlya vuzov, pod red. V.S. Zarubina, A.P. Krishchenko* [Mathematical Statistics: High-school textbook, under the editorship of V.S. Zarubin, A.P. Krishchenko]. Moscow, N.E. Bauman MGTU Publ., 2001. 424 p.

12. Savotchenko S.E., Stukalov V.A. Semanticheskaya mera rezultatov poiskovykh zaprosov [Semantic Measure of Search Query Results]. *Avtomatizatsiya protsessov upravleniya* [Automation of Control Processes], 2013, no. 4 (34), pp. 57–60.