УДК 025.4.036

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ИНФОРМАЦИОННОГО ПОИСКА И ОЦЕНКА ЭФФЕКТИВНОСТИ ПОИСКОВОЙ СИСТЕМЫ

И.В. Тявкин, В.М. Тютюнник

Кафедра «Конструирование радиоэлектронных и микропроцессорных систем», ГОУ ВПО «ТГТУ»

Представлена членом редколлегии профессором Ю.Л. Муромцевым

Ключевые слова и фразы: бинарный поиск; информационный поиск; эффективность поиска.

Аннотация: Приведена модель решения задачи информационного поиска, включающая математическое описание последовательного и бинарного поиска, и модель оценки эффективности поиска, состоящая из точности, полноты, специфичности, избирательности поиска, а также потери информации и поискового шума.

Для удобства поиска информации в сети Интернет программистами мира разработаны поисковые системы, способные осуществлять два вида поиска: обычный и расширенный [1, 3]. Алгоритмы работы поисковых роботов засекречены [4, 6]. Не менее сложной остается задача поиска в локальных информационных системах с большим количеством баз данных [2, 5].

Модель решения задачи информационного поиска представим в виде картежа

$$I = < M_{ij}, Z_\tau, \rho >,$$

где M_{ij} – массив данных ($i = \overline{1, n}$, $j = \overline{1, m}$, n – количество таблиц в j-й базе данных, m – количество баз данных); Z_τ – массив запроса ($\tau = \overline{1, R}$, R – общее количество терминов в запросе); ρ – отношение идентичности, то есть

$$x \rho y \Leftrightarrow x = y.$$

Рассмотрим последовательный поиск. Количество сравнений записей в среднем при удачном поиске составляет $\frac{N+1}{2}$, где N – количество всех записей во всех базах данных.

Если поиск окажется неудачным, то количество сравнений будет равно N.

В бинарном поиске используется бинарное дерево. Для выполнения поиска массив M_{ij} делится на два, в результате чего получается две ветви. Через $\max\limits_{x \in M_{ij}} x$ обозначим такое $x_{\max} \in M_{ij}$, что для любых элементов массива $x \in M_{ij}$ выполнено $x \le x_{\max}$, а через $\min\limits_{x \in M_{ij}} x$ обозначим такое $x_{\min} \in M_{ij}$, что для любых $x \in M_{ij}$ выполнено $x_{\min} \le x$. Тогда

$$g_{\le,y}(x)=\begin{cases}1, \text{ если } x \le y,\\ 2, \text{ в противном случае,}\end{cases} \qquad y \in Z_\tau;$$

$$f_{=,y}(x)=\begin{cases}0, \text{ если } x \ne y,\\ 1, \text{ если } x = y,\end{cases} \qquad y \in Z_\tau;$$

$$G_l = \{g_{\le,y}(x) : y \in Z_\tau\};$$

$$F = \{f_{=,y}(x) : y \in Z_\tau\};$$

$$\Phi_{\text{bin}} = < F, G_l >,$$

где $g_{\le,y}(x)$ – переключатель вершин бинарного дерева; $f_{=,y}(x)$ – предикат ребра бинарного дерева ведущего в лист с записью y; G_l – множество переключателей (функций), определенных на множестве запросов Z_τ и принимающих значения из конечного подмножества натурального ряда; F – множество предикатов, определяемых на множестве запросов Z_τ; Φ_{bin} – базовое множество; y – элемент, принадлежащий множеству запросов Z_τ; Z_τ – массив запроса ($\tau = \overline{1, R}$).

Эффективность поиска (Э) определяется, по крайней мере, двумя основными – точностью и полнотой – и четырьмя дополнительными – специфичностью, избирательностью, коэффициентом потери информации и коэффициентом поискового шума – показателями и имеет вид

$$\text{Э}_k = < \text{Т}_k, \text{П}_k, \text{С}_k, \text{И}_k, \text{ПИ}_k, \text{Ш}_k >,$$

где Т_k – точность поиска ($k = \overline{1, K}$); П_k – полнота поиска; С_k – специфичность поиска; И_k – избирательность поиска; ПИ_k – потери информации; Ш_k – поисковый шум; k – порядковый номер запроса; K – количество запросов.

Для вычисления этих показателей в случае, когда дескриптор запроса пользователя полностью совпадает с найденными в БД данными, используются стандартные формулы [4, 7, 8]

$$\text{П} = \frac{a}{a+c}, \quad \text{Т} = \frac{a}{a+b}, \quad \text{С} = \frac{d}{b+d}, \quad \text{И} = \frac{a+c}{a+c+b+d}, \quad \text{ПИ} = 1 - \text{П}, \quad \text{Ш} = 1 - \text{Т},$$

где a – количество выданных релевантных документов; c – количество релевантных документов в массиве БД, не выданных информационно-поисковой системой (**ИПС**); b – количество выданных ИПС не релевантных документов; d – количество не выданных ИПС не релевантных документов.

Специфичность и избирательность практически применяются при оценке эффективности поиска только в случаях особой необходимости.

Значения a и b определяет пользователь, а значения c и d – эксперт, так как он может выявить как релевантные, так и не релевантные пользовательскому запросу данные в БД. Если поиск выполняется для нахождения идентичных объектов БД, то ИПС всегда выдаст все идентичные записи, соответствующие запросу (то есть П = Т = 1).

В реальных информационных системах полнота поиска по содержанию составляет 60…70 % (0,6…0,7), а точность – 40…50 % (0,4…0,5) [4]. Иногда полнота поиска по содержанию составляет 70…90 % (0,7…0,9), а коэффициент точности обычно находится в пределах 10…100 % (0,1…1,0) [8].

По одному или двум запросам нельзя оценить работу поисковой системы. Для получения более точной оценки проводят K запросов и производят расчет средних значений $\mathrm{T_{ср}}$ и $\mathrm{П_{ср}}$ по следующим формулам:

$$\mathrm{П_{ср}} = \sum_{k=1}^{K} \frac{\mathrm{П}_k}{K}, \quad \mathrm{T_{ср}} = \sum_{k=1}^{K} \frac{\mathrm{T}_k}{K},$$

где $\mathrm{П}_k$ – полнота поиска для k-го запроса; T_k – точность поиска для k-го запроса.

Данный случай оценки эффективности поиска можно использовать только группой экспертов для оценки работы готовой ИПС.

Список литературы

1. Арутюнян, Р.Э. Автоматизация информационного поиска в сети Интернет / Р.Э. Арутюнян // Искусственный интеллект. Интеллектуальные и многопроцессорные системы : материалы Междунар. науч.-техн. конф., Таганрог, 20–25 сент. 2004 г. / Таганрог. гос. радиотехн. ун-т. – Таганрог ; Донецк, 2004. – Т. 1. – С. 353–355.

2. Астанин, С.В. Анализ систем и методов поиска информации в полнотекстовых базах данных / С.В. Астанин // Телекоммуникации и информатизация образования. – 2005. – № 4. – С. 38–45.

3. Воскресенский, А.Л. Формирование запросов к поисковой машине для извлечения знаний из Интернета / А.Л. Воскресенский, Г.К. Хахалин // Компьютерная лингвистика и интеллектуальные технологии : тр. Междунар. конф. «Диалог'2005», Звенигород, 1–6 июня 2005 г. – М., 2005. – С. 86–91.

4. Гусев, В.С. Google : эффективный поиск. Краткое руководство / В.С. Гусев. – М. : Вильямс, 2006. – 240 с.

5. Дикова, Ф.А. Проблема поиска в системе информационного банка данных наукоемких технологий / Ф.А. Дикова, М.В. Куницын // Информация, инновации, инвестиции : материалы междунар. науч.-техн. конф., Уфа, 21–22 нояб. 2007 г. – Уфа, 2007. – С. 47–49.

6. Ефремов, В. Особенности умного поиска / В. Ефремов // Открытые системы. – 2005. – № 11. – С. 48–52.

7. Целых, А.Н. Оценка эффективности информационного поиска / А.Н. Целых, Э.М. Котов // Изв. Таганрог. гос. радиотехн. ун-та. – 2006. – № 10. – С. 43–45.

8. Шемакин, Ю.И. Теоретическая информатика : учеб. пособие / Ю.И. Шемакин ; под ред. К.И. Курбакова. – М. : Изд-во Рос. экон. акад., 1998. – 132 с.

Mathematical Model of Information Search and Estimation of Search System Efficiency

I.V. Tyavkin, V.M. Tyutyunnik

Department «Designing of Radio Electronic and Microprocessor Systems», TSTU

Key words and phrases: binary search; efficiency of search; information search.

Abstract: The paper presents the model of solution to the task of information search including mathematical description of sequential and binary search as well as the model of evaluation of efficiency search including the parameters of accuracy, completeness, specificity and selectivity as well as loss of information and search noise.

Matematisches Modell des Informationssuchens und Einschätzung der Effektivität des Suchsystems

Zusammenfassung: Es ist das Modell der Lösung der Aufgabe des Informationssuchens angeführt. Es schließt die mathematische Beschreibung des aufeinanderfolgenden und binaren Suchens ein. Es ist auch das Modell der Einschätzung der Effektivität des Suchens, das aus der Exaktheit, aus der Fülle, aus der Spezifität, aus der Selektivität des Suchens und auch aus dem Verlust der Information und des Suchenlärms besteht, angeführt.

Modèle mathématique de la recherché informatique et évaluation de l'efficacité du système de recherché

Résumé: Est cité le modèle de la solution du problème de la recherche informatique comprenant la description mathématique de la recherche séquentielle et binaire et le modèle de l'évaluation de l'efficacité de la recherché qui se compose de la précision, de la totalité, de la spécificité, de la sélectivité de la recherché ainsi que de la perte de l'information et du bruit de recherché.