УДК 681.3

**КОМБИНАТОРНЫЙ ПОДХОД К ОПТИМАЛЬНОМУ ПРЕДСТАВЛЕНИЮ ТЕКСТОВЫХ ДОКУМЕНТОВ ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМ**

Занин Дмитрий Евгеньевич
аспирант
*Кубанский государственный технологический университет, Кранодар, Россия*

Задача оптимального представления текстовых документов на заключительном этапе функционирования информационно-поисковой системы представлена как целочисленная, комбинаторная задача о назначениях отранжированному месту в итоговом списке – каждого из найденных документов. Решать задачу предлагается с использованием алгоритма Куна в составе автоматических поисковых серверов.

Ключевые слова: ИНФОРМАЦИОННО-ПОИСКОВАЯ СИСТЕМА, НЕЙРОННАЯ СЕТЬ ХОПФИЛДА, РАНЖИРОВАНИЕ, ОПТИМИЗАЦИЯ.

UDC 681.3

**THE COMBINATORY APPROACH TO OPTIMAL REPRESENTATION OF TEXT DOCUMENTS OF INFORMATION RETRIEVAL SYSTEMS**

Zanin Dmitry Eugenievich
postgraduate student
*Kuban State Technological University, Krasnodar, Russia*

The task of optimal representation of text documents at the final stage of operation of an information retrieval system represented as the integer, combinatory task about assigning to the ranked place in the total list – each of the retrieved documents. To solve the task it is offered with usage of algorithm Kuhn in structure of automatic search servers.

Keywords: INFORMATION RETRIEVAL SYSTEM, NEURAL NETWORK OF HOPFIELD, RANGING, OPTIMIZATION.

Основная проблема, с которой сталкивается любая информационно-поисковая система (ИПС) – как предоставлять пользователю только релевантные документы, согласно его (пользователя) информационных требований, при этом не показывать, или минимизировать число показанных нерелевантных документов.

Располагая несколькими оценками соответствия смыслового содержания найденной информации требованиям поискового запроса, несколькими ключами поиска, задача ранжировки массива документов становится задачей комбинаторной оптимизации большой размерности, решать которую необходимо в реальном масштабе времени.

Наличие нескольких критериев поиска равнозначно наличию у каждого найденного документа нескольких параметров релевантности по каждому из критериев запросов. Таким образом, в общей постановке задача оптимального представления текстовых документов на

заключительном этапе функционирования ИПС выглядит следующим образом:

Дано:

1. Множество $V$ критериев значимости (значимости запросов) информационно-поисковой системы при поиске документов, $V = \{v_i\}$, $i$=1,…,$n$, $n$ – общее количество критериев значимости информационно-поисковой системы на одной итерации ранжирования.

2. Множество $D$ документов найденных ИПС $\{d_s\}$, $s$ = 1,…,$U$, где $U$ – общее количество документов найденных ИПС на одной итерации ранжирования.

3. Каждому элементу $v_i$ множества критериев значимости, в результате статистических исследований либо экспертных оценок, сопоставлена функция цены - $c_i$, $i$=1,...,$N$, характеризующая степень условного "информационного" убытка владельца информационной системы, в случае оценки документа по $i$-му критерию.

4. Каждый документ $d_j$ обладает некоторой релевантностью $g_j$, $j$= 1,…,$M$ на множестве критериев запроса $\{v_i\}$ и стоимостью $p'_j$, $j$= 1,…, $M$ (затратами на его размещение в отранжированном списке результатов поиска ИПС).

Требуется:

Синтезировать алгоритм, позволяющий определить подмножество элементов из множества $\{d_j\}$ документов, при котором реализуется одно из следующих условий:

- минимум суммарного ущерба, при заданной $P$ стоимости (ограничениях на стоимость) размещения документов в отранжированном списке ИПС;

- максимум суммарной релевантности $G$ итогового списка отранжированных документов на множестве $V$ критериев значимости, при заданном $C$ ущербе размещения списка.

Задача, представления найденных ИПС документов – многоитерационная задача поиска оптимального состава отранжированных документов носит целочисленный и комбинаторный характер. Динамическое размещение документов в списке требует оперативного решения многопараметрической задачи в реальном масштабе времени.

Обозначим через $R = \|r_{ji}\|$ - $N \times M$ матрицу производительностей, элементы которой $r_{ji}$ представляют собой релевантность документа с номером $j$ относительно $i$-й позиции в итоговом списке (табл.1).

**Таблица 1 - Таблица задачи о назначениях**

| | Документ 1 | Документ.. | Документ $i$ | Документ... | Документ $N$ |
|---|---|---|---|---|---|
| Позиция 1 | $r_{11}$ | … | $r_{1i}$ | … | $r_{1N}$ |
| Позиция … | … | … | … | … | … |
| Позиция $j$ | $r_{j1}$ | … | $r_{ji}$ | … | … |
| Позиция … | … | … | … | … | … |
| Позиция $M$ | $r_{M1}$ | … | $r_{M\,j}$ | … | $r_{M\,N}$ |

4. Обозначим через $X = \|x_{ji}\|$ $N \times M$ матрицу неизвестных, элемент которой $x_{ji}$ принимает значение 1, если документ с номером $j$ будет находиться в позиции с номером $i$, и значение 0, в противном случае.

5. Ограничения математической модели представлены системой уравнений:

$$\begin{cases} \sum_{j=1}^{M} x_{ji} \le 1,\ i = \overline{1,N}, \\ \sum_{i=1}^{N} x_{ji} \le 1,\ j = \overline{1,M}, \\ x_{ji} \in \{0,1\},\ j = \overline{1,M},\ i = \overline{1,N} \end{cases} \quad (1)$$

Здесь первое уравнения означает, что каждому документу будет назначено не более чем одна (наиболее эффективная) порядковая позиция в итоговом списке.

*Требуется*:

Определить матрицу назначений $X$, при которой имеет место критерий оптимальности:

$$F(X) = \sum_{j=1}^{M} \sum_{i=1}^{N} r_{ji} x_{ji} \to \max . \quad (2)$$

Задача (1) – (2) называется задачей о назначениях с аддитивным критерием оптимальности.

При рассмотрении задачи о назначениях в стандартной форме предполагается, что количество документов равно количеству позиций итогового списка: $M=N$. Нетрудно показать, что введением фиктивных документов или фиктивных номеров позиций математическая модель в открытой форме (1) эквивалентна модели (3).

$$\begin{cases} \sum_{j=1}^{M} x_{ji} = 1, i = \overline{1, N}, \\ \sum_{i=1}^{N} x_{ji} = 1, \ j = \overline{1, M}, \\ 0 \le x_{ji} \le 1, \ j = \overline{1, M}, i = \overline{1, N}, \\ N = M. \end{cases} \quad (3)$$

Исходя из того, что матрица ограничений условий (3) является абсолютно унимодулярной (целочисленная матрица называется абсолютно или вполне унимодулярной, если любой ее минор равен 1, -1 или 0), то любой опорный план математической модели (3) является целочисленным, отсюда вытекает эквивалентность математических моделей (1) и (3) [1]. Кроме того, так как из условий (3) и условий неотрицательности переменных автоматически следует, что переменные не могут быть больше 0, исходная математическая модель (1) эквивалентна (с точки зрения поиска оптимального решения задачи о назначениях) математической модели с ограничениями (3), условиями $M = N$ и ограничениями $x_{ji}, \ge 0$, $j$=1,2,...,$M$, $i$=1,2,...,$N$.

Рассмотрим постановку задачи о назначениях в открытой форме алгоритма Куна (2)-(3). Двойственная к ней задача имеет вид [2]:

$$y_j + z_i \geq r_{ji},\ j = \overline{1,M},\ i = \overline{1,N}, \quad (4)$$

$$Q(y,z) = \sum_{j=1}^{M} y_j + \sum_{i=1}^{N} z_i \to \min,\ j = \overline{1,M},\ i = \overline{1,N}, \quad (5)$$

где $M = N$.

Не уменьшая общности, будем считать, что коэффициенты $r_{ji}$ целые. Пусть $y'$ и $z'$ - допустимое решение задачи (4), (5), т.е. $y'_j + z'_i \geq r_{ji},\ j = \overline{1,M},\ i = \overline{1,N}$.

Допустимое решение может быть построено двумя способами. Пусть $y'_j = \max r_{ji}$, где максимум берется по всем $i$=1,2,...,$N$, $z'_i = 0$, $i$=1,2,...,$N$. Обозначим через $P$ множество тех пар ($j$,$i$), для которых $y'_j + z'_i = r_{ji}$. Рассмотрим простейшую задачу о назначениях с матрицей $D$, элементы которой $d_{ji} = 1$, если $(j,i) \in P$ и $d_{ji} = 0$ в противном случае.

Способ 1. Простейшая задача о назначениях с матрицей $D$ имеет решение, т.е. каждый документ назначается на свою позицию и каждая позиция занимается своим документом. Пусть $X'$ - оптимальное решения простейшей задачи о назначениях, тогда $X'$ - будет оптимальным решением и исходной задачи (2)-(3). Действительно, $x_{ji} = 1$, если $(j,i) \in P$, т.е. $y'_j + z'_{ji} = r_{ji}$, отсюда

$\sum_{j=1}^{M}\sum_{i=1}^{N} r_{ji} x'_{ji} = \sum_{j=1}^{M} y'_j + \sum_{i=1}^{N} z'_i\ \ j = \overline{1,M},\ i = \overline{1,N},$ т.е. по теореме о равенстве линейных форм прямой и двойственной задач , $X'$ - оптимальное решение исходной задачи.

Способ 2. Простейшая задача о назначениях с матрицей $D$ не имеет решения. Тогда найдется множество документов $K$, которые могут назначаться согласно матрице $D$ позициям из множества $Q$, причем мощности множеств равны, соответственно, $k$ и $q$ и при этом $k<q$. Рассмотрим новые двойственные переменные:

$y''_j = y'_j - 1$ , если $j \in K$ и $y''_j = y'_j$ в противном случае;

$z''_i = z'_i + 1$ , если $i \in Q$ и $z''_i = z'_i$ в противном случае.

Новые значения двойственных переменных удовлетворяют условиям задач (4), (5) и при этом уменьшают значения критерия двойственной задачи.

Переходим на начало процедуры решения задачи и так до тех пор, пока на очередном шаге не получим решение простейшей задачи о назначениях, которое и определит оптимальное решение исходной задачи.

Конечность алгоритма Куна следует из того, что по теореме о соотношениях линейных форм прямой и двойственной задач $\sum_{j=1}^{M}\sum_{i=1}^{N} r_{ji}\, x'_{ji} \le \sum_{j=1}^{M} y'_j + \sum_{i=1}^{N} z'_i \;\; j=\overline{1,M},\; i=\overline{1,N}$.

## Литература


1. Гэри М., Джонсон Д. Вычислительные машины и труднорешаемые задачи.- М.: Мир, 1982. – 416с.
2. Таха Х. Введение в исследование операций. Т. 1. - М.: Мир, 1985.- 282с.
3. Ловас Л., Пламмер М. Прикладные задачи теории графов. Теория паросочетаний в математике, физике, химии. - М.: Мир, 1998.