УДК 005; 007; 681.5

М.Ю. Фролов

Харьковский национальный университет радиоэлектроники, Харьков

МОДЕЛЬ ОРГАНИЗАЦИИ ИССЛЕДОВАТЕЛЬСКОГО ПОИСКА В ЭЛЕКТРОННЫХ АРХИВАХ НА ОСНОВЕ ЕСТЕСТВЕННОЙ КЛАССИФИКАЦИИ

Проанализирована существующая основная модель поиска документов электронных архивов. Предложена новая организация исследовательского поиска в электронных архивах на основе естественной классификации и ее модель.

Ключевые слова: *электронный архив, поиск, индексная база, естественная классификация, граф, документ.*

Введение

Электронный архив – неотъемлемая часть современного электронного документооборота. Современные электронные архивы содержат не только образы бумажных документов и электронные документы. Архивируются самые разные документы и данные (например, корпоративная электронная почта). Во многих странах наличие архива электронной почты уже стало законодательным требованием, предусматривается хранение почтовых сообщений от 7 до 26 лет. Несоблюдение этих условий грозит организациям выплатой огромных штрафов и даже уголовным преследованием [1, 2]. Одной из основных и наиболее востребованной функций электронного архива (ЭА) является быстрый и эффективный поиск.

Существует два подхода к поиску документов. Первый подход состоит в том, что в процессе поиска ищется документ, который точно существует в системе, и задача состоит в том, чтобы процесс поиска свелся к нахождению требуемого документа или документов – точный поиск. Второй подход состоит в том, что ищутся все документы по интересующему вопросу – исследовательский поиск [3]. Для последнего подхода присуще такое понятие, как шум – часть документов в результирующем наборе, где встретились подобные по синтаксису, но не по смыслу слова [3], т.е. документов нерелевантных поисковому запросу. В работе [4] описаны и проанализированы некоторые проблемы поиска в ЭА, в том числе исследовательском поиске, а также предложен новый подход в осуществлении исследовательского поиска.

Формулировка цели статьи. Целью данной статьи является построение математической модели, которая позволит снизить уровень шума в результирующем наборе документов, организации исследовательского поиска в электронных архивах на основе естественной классификации.

Изложение основного материала

Рассмотрим основной метод поиска – на базе инвертированной матрицы, где словам или нормализованным словоформам ставятся в соответствие адреса документов. Суть подхода заключается в том, что при создании индексных файлов (индексированной матрицы или индексной базы данных) в них вносятся все значимые слова (без союзов, предлогов и т.п.) из всех документов в алфавитном порядке. Эти слова затем объединяются в пары с указателями на документы, содержащие эти слова.

Пусть Г1 (рис. 1) – граф организации поиска в электронном архиве. Граф Г1 состоит из: множества вершин $V^1(Г1) = \{V^1_i\}$, где $i = \overline{1, n+t}$, V^1_0 – индексная база данных, $V^1_1 - V^1_n$ - индексы, где n – конечное количество индексов; $V^1_{n+1} - V^1_{n+t}$ – документы, где n+t – конечное количество документов; ребер $E^1(Г1) = \bigcup_{k=1}^{2} E^1_k$, где $E^1_i \bigcap E^1_j = \varnothing$ при $i \neq j$, $E^1_1 = \{e^1_1, e^1_2, e^1_3, e^1_4, \dots, e^1_n\}$ – «указатель на индекс», где n – конечное количество индексов; $E^1_2 = \{e^1_{n+1} - e^1_{n+m}\}$ – «ссылка на документ», где m – конечное количество ссылок на документы.

Индекс ($И_1, И_2, И_3, И_4 \dots И_n$) представляет собой слово, содержащееся в документах ($Д_1, Д_2, Д_3, Д_4, Д_5, Д_6 \dots Д_n$), хранимых в электронном архиве, а также ссылку непосредственно на документ, содержащий это слово. При осуществлении поиска по электронному архиву поисковая система ЭА просматривает индексную базу данных, находя слова из запроса пользователя, и предоставляет все документы, в которых содержатся данные слова. В результате мы сталкиваемся с проблемой релевантности найденных документов поисковому запросу.

Для минимизации документов в результирующем наборе, где встретились подобные по синтаксису, но не по смыслу слова, предлагается использовать параметрическую классификацию [5].

Пусть Г2 (рис. 2) – граф классификации понятий предметной области электронного архива. Граф Г2 состоит из: множества вершин $V^2(Г2)=\{V^2_i\}$, где $i=\overline{1, n+t}$, V^2_0 – классификация понятий, $V^2_1 - V^2_n$ – понятия ($П_1, П_2, П_3, \dots, П_n$), где n – конечное количество понятий; $V^2_{n+1} - V^2_{n+t}$ – конечные объекты –

документы ($Д_1$,$Д_2$,$Д_3$,$Д_4$,$Д_5$, …,$Д_n$), где n+t – конечное количество документов; ребер $E^2(Г2)= \{E_i^2\}$, где $i=\overline{1,n+m}$ – отношения. Классификация понятий является частью параметрической классификации.

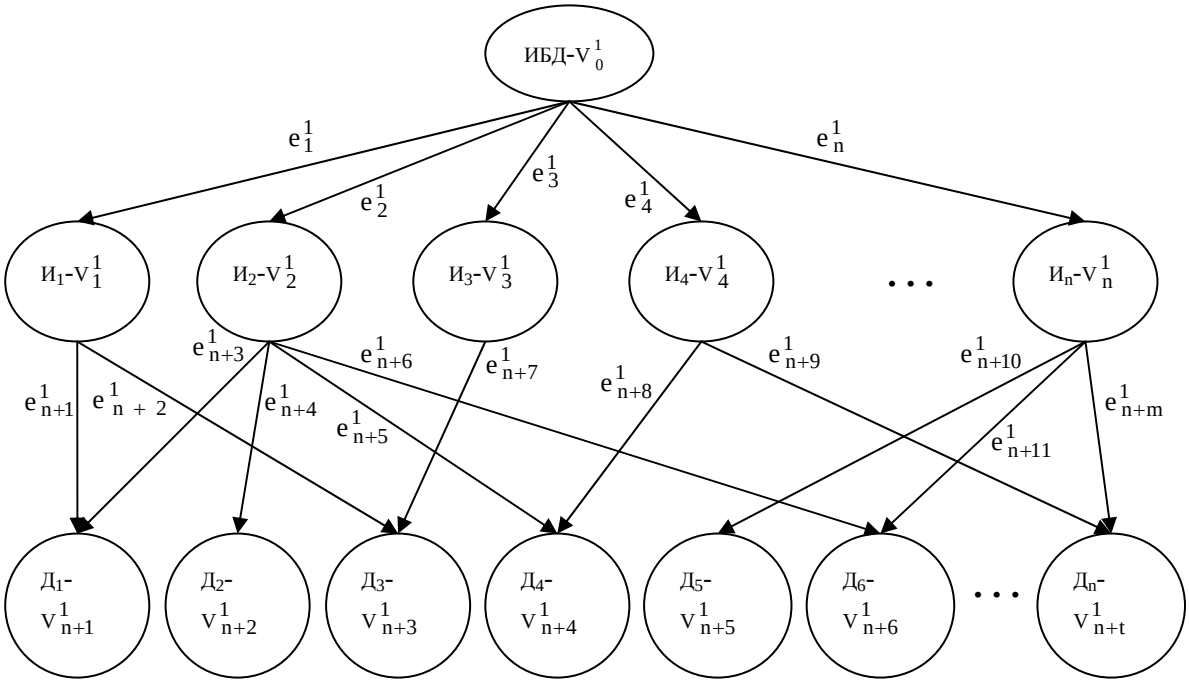

Рис. 1. Граф организации поиска в электронном архиве

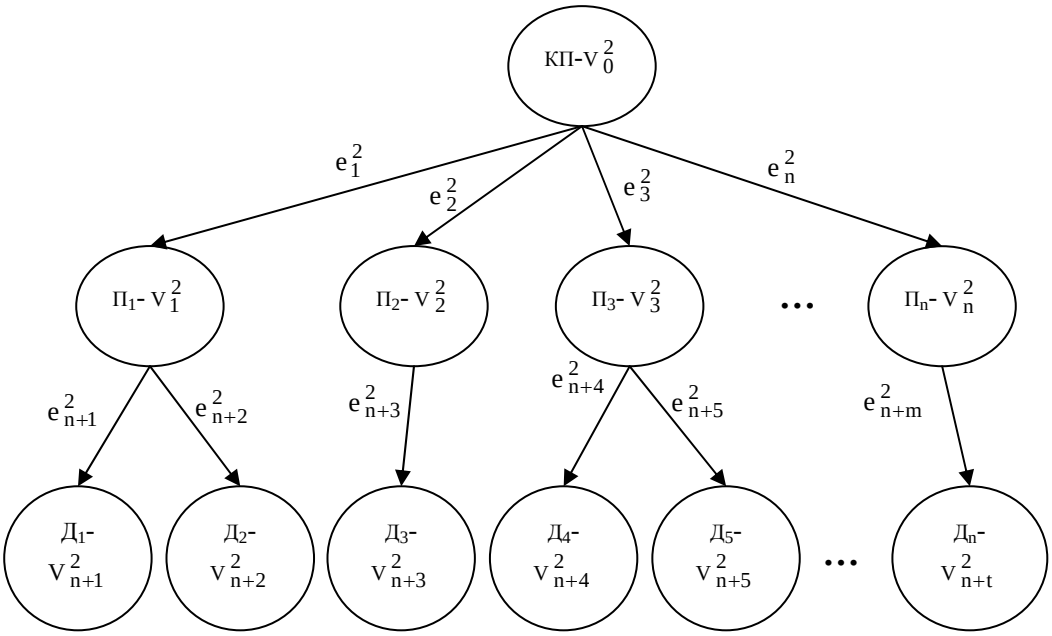

Рис. 2. Граф фрагмента классификации понятий

У каждого объекта (вид) может быть только один предок (надвид), т.е. узел. У каждого элемента классификации объектов имеется главное функциональное свойство. Упорядоченная, в соответствии с классификацией понятий, совокупность функциональных свойств называется классификацией свойств.

Общий вид параметрической классификации представлен ниже (рис. 3). Пусть ГЗ (рис. 3) – граф параметрической классификации. Граф ГЗ состоит из: множества вершин $V^3(ГЗ)=\{V_i^3\}$, где $i=\overline{1,9}$, V_i^3 – параметрическая классификация (ПК), V_2^3 – классификация понятий (КП), V_3^3 – классификация свойств (КС), V_4^3 – понятие (П), V_5^3, V_6^3 – конечный объект (O_1, O_2), V_7^3 – V_9^3 – функциональные свойства ($ФС_1$,$ФС_2$,$ФС_3$); ребер $E^3(ГЗ)= \{E_i^3\}$, где $i=\overline{1,11}$ – отношения. Для улучшения эффективности поиска (снижения шума) предлагается создать, помимо основной индексной базы данных (рис. 1), дополнительную индексную базу. Дополнительная индексная база данных должна содержать индексные файлы,

которые включают в себя слова и словосочетания из функциональных свойств классификации свойств в процессе индексации. Именно по дополнительной базе будет проводиться первичный поиск документов, т.е. поиск будет выглядеть следующим образом (рис. 4).

Пусть Г4 (рис. 4) – граф организации исследовательского поиска на основе естественной классификации. Граф Г4 состоит из: множества вершин $V^4(Г4)=\{V_i^4\}$, где $i=\overline{1,n+m}$, V_1^4 – дополнительная индексная база данных (ДИБД), V_2^4, V_5^4, … ,V_{n-2}^4 – индексы ($И_1$, $И_{2,\dots,}$ $И_n$, где n – конечное количество индексов), V_3^4, V_6^4, … , V_{n-1}^4 – функциональные свойства ($ФС_1$, $ФС_2$, … ,$ФС_n$, где n – конечное количество функциональных свойств), V_4^4, V_7^4, … , V_n^4 – понятия ($П_1$, $П_2$, … ,$П_n$), V_{n+1}^4, V_{n+2}^4, V_{n+3}^4, … , V_{n+m}^4 – документы ($Д_1$, $Д_2$, $Д_3$, … , $Д_n$, где n – конечное количество документов); ребер $E^4(Г4)=\{E_i^4\}$, где $i=\overline{1,t}$ – ссылки.

Граф Г4 – пересечение графа Г1 и графа Г2 с учетом классификации свойств. Пересечение графов G_1 и G_2, обозначаемое как $G_1 \cap G_2$, представляет собой граф $G_4 = (X_1 \cap X_2, A_1 \cap A_2)$. Таким образом, множество вершин графа G_4 состоит из вершин и ребер, присутствующих одновременно в G_1 и G_2 [1]. Рассмотрим пересечение графов Г1 и Г2, здесь V_0^1 и V_0^2 совпадут; так как множество документов неизменно, то $|\{V_i^1\}|=|\{V_j^2\}|=n$, $\{V_i^1\}$ и $\{V_j^2\}$ также совпадут, где n – конечное множество документов, $i=\overline{6,12}$, $j=\overline{5,10}$; вершины $V_1^1 - V_n^1$ и $V_1^2 - V_n^2$ совпадут, а также $e_{n+1}^1 - e_{n+m}^1$ и $e_{n+1}^2 - e_{n+m}^2$ совпадут таким образом, что конечных объектов (документов) графа Г4, т.е. вершины $V_{n+1}^4 - V_{n+m}^4$ будут иметь только один элемент верхнего уровня (предок). Однако граф Г4 неполный, так как не учитывает связь «понятие»–«функциональное свойство».

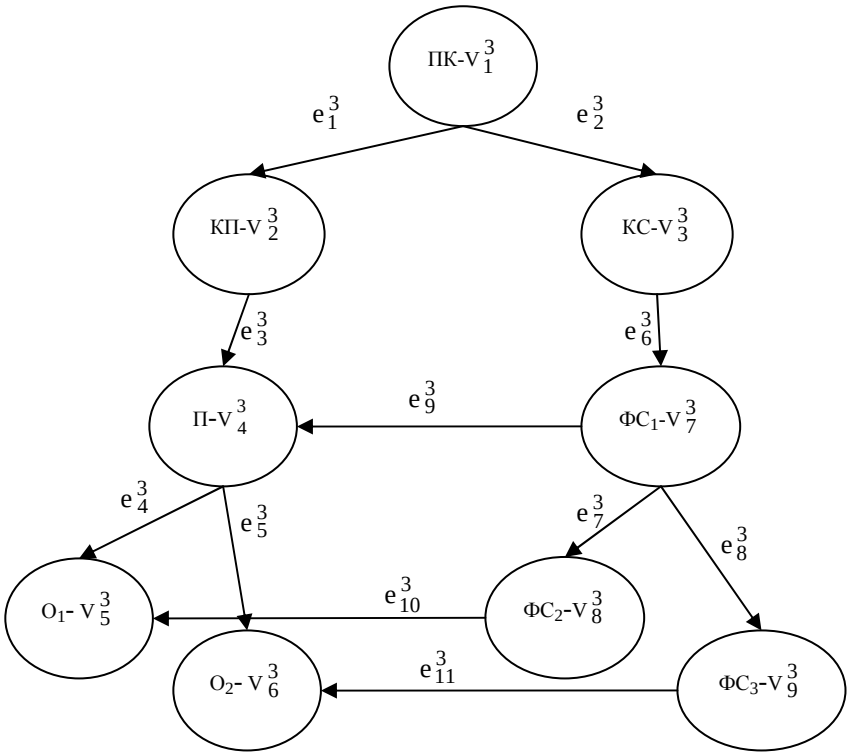

Рис. 3. Граф общего вида параметрической классификации

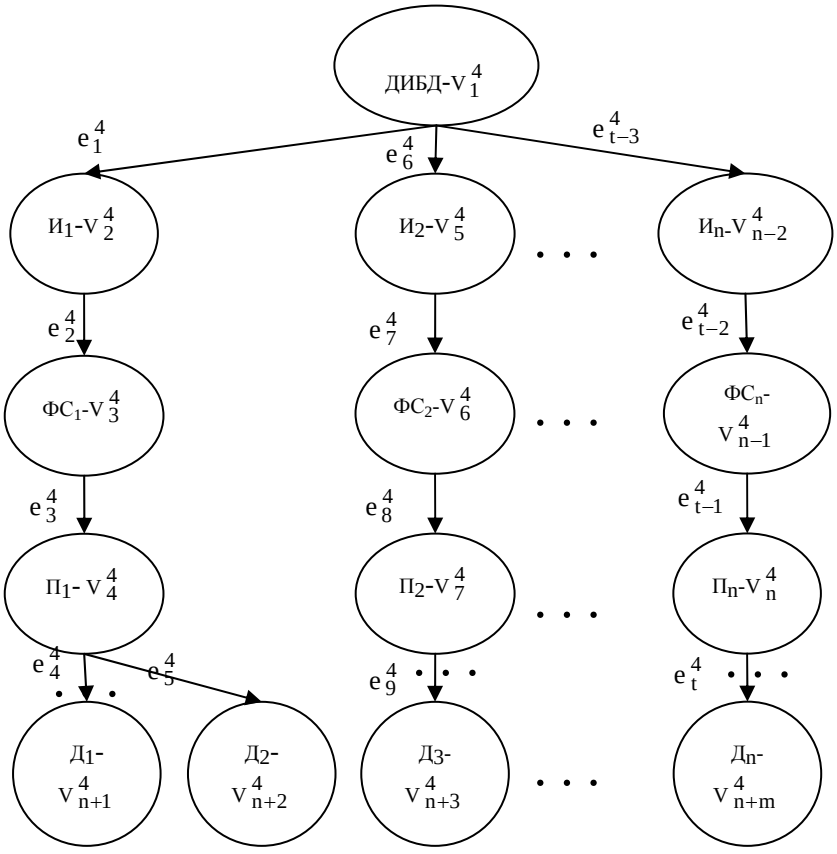

Рис. 4. Граф организации исследовательского поиска на основе естественной классификации

Поэтому в графе Г4 добавлены вершины V_3^4, V_6^4, V_{n-1}^4. Матрица смежности для графа Г4 будет выглядеть следующим образом (табл. 1).

Таблица 1

Матрица смежности графа Г4

	V_1^4	V_2^4	V_3^4	V_4^4	V_5^4	V_6^4	V_7^4	V_{n-2}^4	V_{n-1}^4	V_n^4	V_{n+1}^4	V_{n+2}^4	V_{n+3}^4	V_{n+m}^4
V_1^4		1			1			1						
V_2^4			1											
V_3^4				1										
V_4^4											1	1		
V_5^4						1								
V_6^4							1							
V_7^4													1	
V_{n-2}^4									1					
V_{n-1}^4										1				
V_n^4														1
V_{n+1}^4														
V_{n+2}^4														
V_{n+3}^4														
V_{n+m}^4														

В результате поиска по дополнительной индексной базе будет найдено функциональное свойство, которое наиболее совпадает с параметрами поискового запроса. Этому свойству сопоставляется соответствующее понятие из классификации понятий. Затем находится ветвь (ветви) классификации понятий, содержащая все документы, удовлетворяющие запросу пользователя.

Отсюда следует, что при поиске будут отсеиваться документы, нерелевантные запросу. Для увеличения эффективности результатов поиска можно проводить поиск документов, используя основную индексную базу данных среди найденных документов при первичном поиске.

Выводы

Разработана модель организации исследовательского поиска документов в электронных архивах на основе естественной классификации. Разработанная модель позволяет снизить уровень шума в результирующем наборе документов за счет построения новой (дополнительной) индексной базы данных, которая учитывает функциональные свойства объектов ЭА.

Список литературы

1. Фролов М.Ю. Обзор и анализ основных систем автоматизации документооборота / М.Ю. Фролов // Системи обробки інформації. – Х.: ХУ ПС, 2009. – Вип. 3(77). – С. 131-134.

2. Хахамов С. Электронный архив как ядро информационно-коммуникационной системы предприятия / Сергей Хахамов [Электронный ресурс]. – Режим доступа к ресурсу: http://www.connect.ru/article.asp?id=6260.

3. Гавердовский А. Концепция построения систем автоматизации документооборота / Анатолий Гавердовский // Открытые системы. – АО ВЕСТЬ. – 1997. – № 1. – С. 29-34.

4. Соловьева Е.А. Организация логической структуры электронных архивов на основе естественной классификации / Е.А. Соловьева, М.Ю. Фролов // Вестник ХПИ. – 2009. – С. 183-188.

5. Соловьева Е.А. Естественная классификация: системологические основания / Е.А. Соловьева. – Х.: ХНУРЕ, 1999. – 222 с.

Поступила в редколлегию 2.04.2010

Рецензент: д-р техн. наук, проф. Е.А. Соловьева, Харьковский национальный университет радиоэлектроники, Харьков.

МОДЕЛЬ ОРГАНИЗАЦІЇ ДОСЛІДНИЦЬКОГО ПОШУКУ В ЕЛЕКТРОННИХ АРХІВАХ НА ОСНОВІ ПРИРОДНЬОЇ КЛАСИФІКАЦІЇ

М.Ю. Фролов

Проаналізовано існуючу основну модель пошуку документів електронних архівів. Запропоновано нову організацію дослідницького пошуку в електронних архівах на основі природної класифікації та її модель.

Ключові слова: *електронний архів, пошук, індексна база, природна класифікація, граф, документ.*

MODEL OF THE RESEARCH SEARCH IN THE ELECTRONIC ARCHIVE BASED ON NATURAL CLASSIFICATION

M.Y. Frolov

We analyzed the existing basic model document search of electronic archives. A new organization of research in search of electronic archives on the basis of natural classification and its model.

Keywords: *electronic archive, search, index base, natural classification, graph, paper.*