УДК 004.656

05.00.00 Технические науки

ОРГАНИЗАЦИЯ СЛОВАРЯ ПОНЯТИЙ СИСТЕМЫ ИДЕНТИФИКАЦИИ СВЕДЕНИЙ ОГРАНИЧЕННОГО РАСПРОСТРАНЕНИЯ

Харитоненко Виталий Григорьевич
Адъюнкт очной штатной адъюнктуры
Краснодарское высшее военное училище, Краснодар, Россия
E-mail (trafik2006@yandex.ru)
Телефон: 8-919-195-12-44.

В статье рассматривается задача повышения эффективности контроля информации ограниченного распространения, циркулирующей в информационных сетях общего пользования, посредством разработки системы автоматизированной идентификации сведений ограниченного распространения. Целью разработки данной системы является своевременное предотвращение и идентификация информации ограниченного распространения. Предлагается методика построения словаря как одного из этапов разработки системы автоматизированной идентификации сведений ограниченного распространения. Описывается: 1) порядок объединения словоформ, имеющих одно семантическое значение в понятия, которые обозначаются простыми числами; 2) порядок объединения понятий в сведения, обозначаемые натуральными числами, при этом идентификатор сведения является произведением идентификаторов понятий; 3) порядок представления словоформ их графическими основами; 4) назначение внутрифразовых соединителей в понятиях представленных словосочетаниями; 5) порядок представления словоформ в виде абстрактного типа данных − префиксного дерева; 6) порядок объединения понятий формализуемого документа в общее префиксное дерево графических основ словоформ; 6) порядок идентификации графических основ в дереве. Определяются: 1) ограничения на размер словаря при использовании 64-разрядных процессоров, в случае если операционная система не поддерживает арифметику многократной точности; 2) максимальное количество понятий в одном анализируемом фрагменте текста; 3) максимальное значение идентификатора понятий и максимальное количество понятий в словаре. Представлена таблица, иллюстрирующая зависимость между этими тремя величинами

Ключевые слова: СЛОВАРЬ, ПРЕФИКСНОЕ

UDC 004.656

Technical sciences

ORGANIZATION OF A CONCEPT DICTIONARY FOR IDENTIFICATION SYSTEM OF DATA WITH LIMITED DISTRIBUTION

Haritonenko Vitaly Grigorevich
postgraduate student of internal regular
Krasnodar higher military school, Krasnodar, Russia
E-mail (trafik2006@yandex.ru)
Phone: 8-918-195-12-44

The article deals with the problem of efficiency increase of the control of the information with limited distribution, which circulates in general purpose information networks, by means of working out an automated identification system of data with limited distribution. The purpose of working out the system is timely identification and prevention of leakage of information with limited distribution. There is a technique to construct a dictionary as a phase of working out an automated identification system of data with limited distribution suggested. It describes: 1) an order of association of the word forms having one semantic value in concepts which are designated by simple numbers; 2) the order of association of concepts in the data designated by natural numbers, thus the data identifier is a product of concept identifiers; 3) an order of representation of word forms their graphic bases; 4) designation of intraphrase connectors in concepts presented by word-combinations; 5) an order of representation of word forms in the form of abstract type of data - prefix tree; 6) an order of association of concepts of the formalizable document in the general prefix tree of graphic bases of word forms; 7) an order of identification of graphic bases in the tree. It defines: 1) restrictions on the size of the dictionary at the use of 64-digit processors, in a case if the operational system does not support arithmetics of repeated accuracy; 2) a maximum quantity of concepts of an analyzed fragment of the text; 3) the maximum value of the identifier of concepts and a maximum quantity of concepts of the dictionary. There is a table presented, illustrating the correlation between these three values

Keywords: DICTIONARY, PREFIX TREE, CONCEPT,

ДЕРЕВО, ПОНЯТИЕ, ИДЕНТИФИКАТОР ПОНЯТИЯ, СВЕДЕНИЕ, МНОЖЕСТВО ПОНЯТИЙ, ИДЕНТИФИКАТОР СВЕДЕНИЯ, ГРАФИЧЕСКАЯ ОСНОВА, ВНУТРИФРАЗОВЫЙ СОЕДИНИТЕЛЬ, ФЛЕКСИЯ, ЛОЖНОЕ СРАБАТЫВАНИЕ, ИДЕНТИФИКАТОР ОКОНЧАНИЯ ПОИСКА

CONCEPT IDENTIFIER, DATA, SET OF CONCEPTS, DATA IDENTIFIER, GRAPHIC BASIS, INTRAPHRASE CONNECTOR, INFLECTION, FALSE OPERATION, IDENTIFIER OF THE TERMINATION OF SEARCH

Введение

В соответствии с федеральным законом «Об информации, информационных технологиях и о защите информации» [5] информация в зависимости от категории доступа к ней подразделяется на общедоступную информацию, а также на информацию, доступ к которой ограничен федеральными законами (информация ограниченного доступа). Ограничение доступа к информации устанавливается федеральными законами в целях защиты основ конституционного строя, нравственности, здоровья, прав и законных интересов других лиц, обеспечения обороны страны и безопасности государства. Так же законодательство определяет иные виды тайны, необходимость соблюдения конфиденциальности информации и ответственность за ее разглашение. Согласно Кодекса Российской Федерации об административных правонарушениях [4] разглашение информации, доступ к которой ограничен федеральным законом, влечет наложение административного наказания.

В связи с этим задача разработки интеллектуальных систем защиты информации, предназначенных для контроля и предотвращения утечки информации ограниченного распространения, становится всё более актуальной. Существующие средства поиска, представленные большим количеством программам-поисковиков, осуществляют поиск по ключевых словам. Удаление искомого ключевого слова в проверяемом документе приводит к тому, что программа, например поисковая утилита AVSearch, не обнаружит информацию ограниченного распространения. Персональная поисковая система Архивариус 3000 позволяет восстановить удаленное

слово, однако для передачи смысла всего сообщения одного слова не достаточно, необходимо несколько слов, характеризующих искомое явление, объект, процесс и так далее. Перечисленные средства поиска информации в тексте работают по принципу поиска слов (словосочетаний), задаваемых оператором. При этом оператору приходится знать общую тематику информации, которую необходимо найти в тексте, чтобы задавать соответствующие поисковые образы. Предлагаемый подход отличается от описанного: слова, все их синонимы, близкие по значению слова объединяются в понятия, понятия, описывающие конкретное смысловое значение объединяются в сведения, которые затем включаются в общий словарь. Тем самым исключается необходимость помнить все поисковые образы, так как они собраны в словаре. Система в автоматизированном режиме осуществляет поиск в тексте необходимых сведений.

Программные средства, идентифицирующие информацию ограниченного распространения в автоматизированном режиме, позволят:

выполнять предварительную оценку информации;

уменьшить время работы должностных лиц;

повысить в целом уровень защищенности информации ограниченного распространения, и как следствие – основ конституционного строя, нравственности, здоровья, прав и законных интересов граждан, обеспечения обороны страны и безопасности государства.

Словарь

Словарь автоматизированной системы идентификации сведений ограниченного распространения реализуется с помощью *префиксного дерева* [1]. В вершинах дерева хранятся префиксы слов. Над деревом производятся три основные операции: добавление, удаление и поиск слова.

Операция добавления выполняется на стадии формирования словаря в процессе формализации информации из перечня сведений конфиденциального характера [6]. Известно [6] семь пунктов, которые определяют сведения конфиденциального характера. В первом пункте перечня указаны персональные данные. Отношения, связанные с обработкой персональных данных регулирует федеральный закон «О персональных данных» [7].

Операция удаления может выполняться при модификации словаря в случае внесения изменений в перечень сведений конфиденциального характера.

Поиск слова − это основная операция, которая выполняется при работе автоматизированной системы. При поиске происходит последовательное сравнение символов слов предложения с символами словаря.

Слова, их синонимы, словосочетания, всевозможные сокращения (в зависимости от частоты употребления в специальном лексиконе, решения экспертов [3]), имеющие семантически близкие значения, при составлении словаря объединяются в *понятия*. Например, понятие «покупает» выражается словоформами: *покупает, приобретает, делает покупки*.

Каждому понятию ставится в соответствие простое число, которое является *идентификатором понятия* λ. Например, понятие «покупает» может иметь идентификатор понятия $\lambda = 2$.

Сведение (сведение ограниченного распространения) − это расположенное в пределах одного предложения множество понятий, имеющее конкретное смысловое значение, которое совпадает со значением одного из семи пунктов [6]. Мощность множества понятий всегда больше единицы.

Формализовать представление сведения в виде множества понятий возможно, применив основную теорему арифметики [2]. Пусть

натуральное $N > 1$ является идентификатором сведения. Тогда простые множители, на которые оно разлагается, являются идентификаторами понятий λ. Единственность разложения натурального числа на простые множители соответствует единственному представлению идентификатора сведения в виде множества идентификаторов понятий. Сведения могут выражаться различными множествами понятий и, следовательно, иметь ряд идентификаторов $N_1, N_2, ...N_n$.

Словоформы каждого понятия представляются в виде *графических основ*, которые объединяются в префиксное дерево. В случае, когда понятия выражаются словосочетанием, между словоформами словосочетания в префиксном дереве располагаются *внутрифразовые соединители*, обозначенные шестиугольником (рис. 1). Их назначение – пропуск флексии предыдущего слова и всех символов-сепараторов вплоть до начала следующего слова в словосочетании. Корень дерева обозначается пустой вершиной. Использование графических основ вместо словоформ делает систему безразличной к склонениям слов, что повышает скорость поиска и устойчивость к грамматическим ошибкам, но и повышает вероятность ложного срабатывания. Однако статистические измерения показывают, что для официально-делового стиля документов и узкой предметной области эта вероятность стремится к нулю.

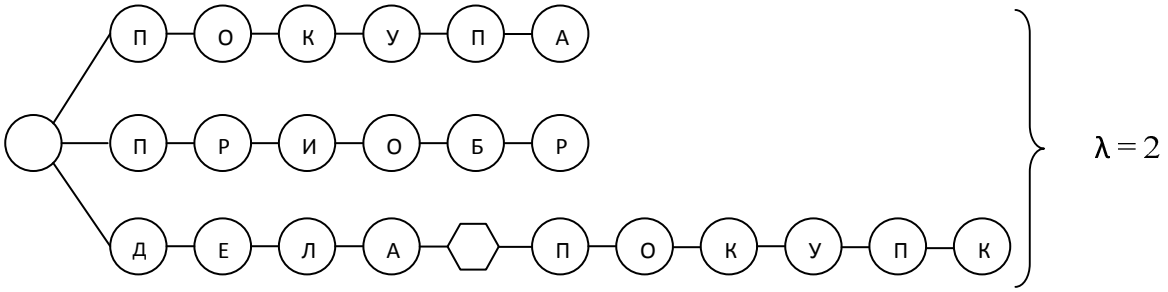

Рисунок 1. Представление понятия «покупает» в виде префиксного дерева

Графические основы словоформ всех понятий добавляются в дерево. Последний символ каждой графической основы снабжается идентификатором понятия, который совпадает индикатором успешного

нахождения понятия в проверяемом фрагменте текста. На рисунке 2 представлено три отдельных понятия: «покупать» (*покупать, приобретать, делать покупки*), «родился» (*родился, появился на свет*) и «улица» (*улица, проспект, переулок*).

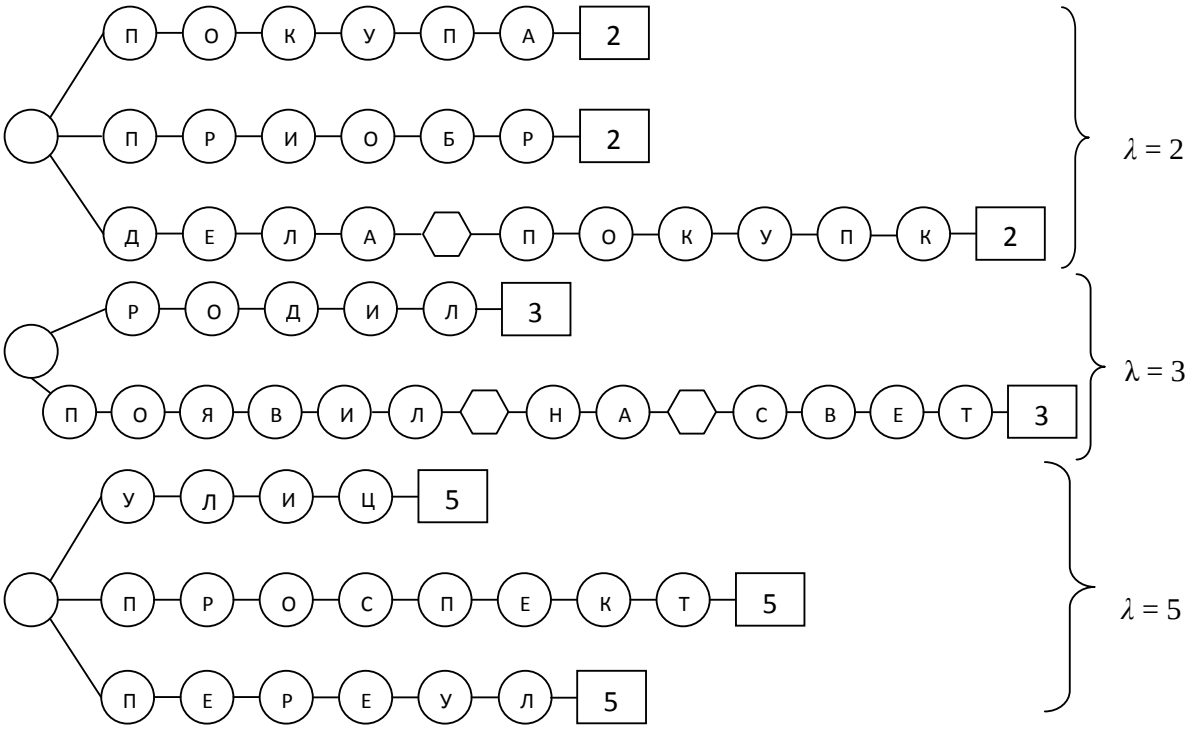

Рисунок 2. Три понятия

На заключительном этапе построения словаря все понятия объединяются в общее префиксное дерево графических основ словоформ. Дерево имеет единственный корень, обозначенный пустой вершиной (рис. 3). Его предназначение – определение начала поиска словоформ из дерева в проверяемом фрагменте текста. Одинаковые префиксы в различных словоформах префиксного дерева объединяются. Применение префиксного дерева позволяет осуществлять параллельный поиск всех словоформ словаря одновременно за минимальное время. Совмещение позиции поиска с идентификатором окончания поиска, обозначенным прямоугольником (рис. 3), позволяет сделать вывод об окончании поиска словоформы из общего дерева во фрагменте текста. Понятие

идентифицируется по идентификатору понятия, совпадающему с идентификатором окончания поиска (рис. 3).

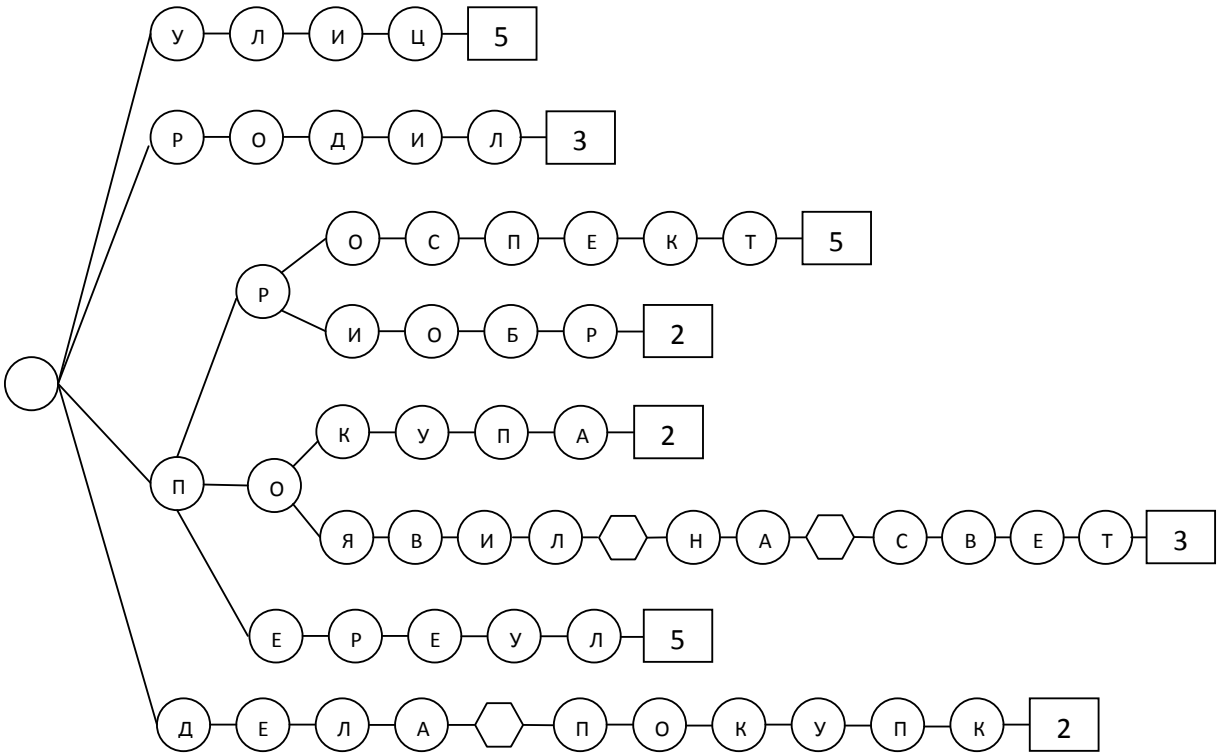

Рисунок 3. Словарь трёх понятий

Ограничения на размер словаря

Процессоры современных компьютеров имеют 64-разрядную архитектуру. Это означает, что они могут оперировать числами не более чем $2^{64}-1$, если имеющаяся программа не обеспечивает поддержку арифметики многократной точности. Это число накладывает косвенные ограничения на максимальное количество понятий, которые могут находиться в одном фрагменте текста, в котором происходит поиск сведений ограниченного распространения. Максимальное количество понятий t в анализируемом фрагменте текста определяет максимальное значение идентификатора понятия $d_{\max} = \max(\lambda)$ и количество понятий в

словаре $\pi(d_{\max})$. Используя таблицу простых чисел, можно подсчитать зависимость этих трех величин при использовании 64-разрядной архитектуры.

Величина $d_{\max}$ рассчитывается следующим образом. Для случая, когда сведения выражаются не более чем тремя понятиями в предложении $d_{\max} = 2642257$. Для случая, когда сведения выражаются не более чем четырмя понятиями в предложении $d_{\max} = 65543$. Идентификаторы понятий λ подбираются таким образом, что их произведение для идентификации одного сведения не превышало значения $2^{64} - 1$.

t	d_{max}	$\pi(d_{max})$
3	2642257	192723
4	65543	6541
5	7151	910
6	1637	253
7	577	99
8	271	50

Таблица 1. Зависимость между t и $d_{\max}$

Анализ текстов, содержащих информацию ограниченного распространения, показывает, что сведение (сведение ограниченного распространения) выражается не менее чем тремя понятиями. В случае $t = 3$, при использовании 64-разрядного процессора, словарь может содержать не более 192723 понятий. В случае $t = 4$, словарь может содержать не более 6541 понятий. И так далее, согласно таблицы 1, иллюстрирующей зависимость между t, $d_{\max}$ и $\pi(d_{\max})$.

Составляя словарь автоматизированной системы идентификации сведений ограниченного распространения, чаще встречающимся понятиям необходимо присваивать идентификатор λ меньшей величины, реже встречающимся понятиям – большей величины. Так же необходимо

принимать в учёт информацию, представленную в таблице 1, то есть в случае если сведение включает три понятия, их величины не должны превышать значений: 2642231, 2642239 и 2642257; если сведение включает четыре понятия, их величины не должны превышать значений: 65521, 65537, 65539 и 65543 соответственно, согласно таблицы 1.

Заключение

В статье описан порядок построения словаря системы автоматизированной идентификации сведений ограниченного распространения. Обозначение понятий простыми числами используется для того, что бы искать сведения ограниченного распространения с использованием основной теоремы арифметики. Основная теорема арифметики так же используется для того, чтобы представить сведение (сведения, составляющие государственную тайну) в виде множества понятий единственным образом. Определены ограничения на размер словаря.

Список литературы:

1. Ахо, А.В., Хопкрофт, Д.В., Ульман, Д.Д., Структуры данных и алгоритмы. Перевод с английского. – М.: Издательский дом Вильямс, 2000. – 383 с.

2. Бухштаб А.А. Теория чисел. – М.: Просвещение, 1965. – 384 с.

3. Гаврилова, Т.А., Хорошевский, В.Ф. Базы знаний интеллектуальных систем. – Санкт-Петербург: Изд-во Питер, 2001. – 384 с.

4. Кодекс Российской Федерации об административных правонарушениях: [введен в действие федеральным законом от 30 декабря 2001 года № 195-ФЗ]−режим доступа: http://www.consultant.ru/popular/koap/ (дата обращения: 25.07.2015).

5. Об информации, информационных технологиях и о защите информации: федер. закон ФЗ-149: [подп. Президентом Российской Федерации 27 июля 2006 г.]. − режим доступа: http://www.rg.ru/2006/07/29/informacia-dok.html (дата обращения: 25.07.2015).

6. Об утверждении перечня сведений конфиденциального характера: [утверждён Указом Президента Российской Федерации от 6 марта 1997 г. № 188] режим доступа: http://base.consultant.ru/cons/cgi/online.cgi?req=doc;base=LAW;n=182734 (дата обращения: 25.07.2015).

7. О персональных данных: федер. закон ФЗ-152 [подп. Президентом Российской Федерации 27 июля 2006 г.] режим доступа: http://base.garant.ru/10200083/#friends (дата обращения: 25.07.2015).

References:

1. Aho, A.V., Hopkroft, D.V., Ul'man, D.D., Struktury dannyh i algoritmy. Perevod s anglijskogo. – M.: Izdatel'skij dom Vil'jams, 2000. – 383 s.

2. Buhshtab A.A. Teorija chisel. – M.: Prosveshhenie, 1965. – 384 s.

3. Gavrilova, T.A., Horoshevskij, V.F. Bazy znanij intellektual'nyh sistem. – Sankt-Peterburg: Izd-vo Piter, 2001. – 384 s.

4. Kodeks Rossijskoj Federacii ob administrativnyh pravonarushenijah: [vveden v dejstvie federal'nym zakonom ot 30 dekabrja 2001 goda № 195-FZ]− rezhim dostupa: http://www.consultant.ru/popular/koap/ (data obrashhenija: 25.07.2015).

5. Ob informacii, informacionnyh tehnologijah i o zashhite informacii: feder. zakon FZ-149: [podp. Prezidentom Rossijskoj Federacii 27 ijulja 2006 g.]. − rezhim dostupa: http://www.rg.ru/2006/07/29/informacia-dok.html (data obrashhenija: 25.07.2015).

6. Ob utverzhdenii perechnja svedenij konfidencial'nogo haraktera: [utverzhdjon Ukazom Prezidenta Rossijskoj Federacii ot 6 marta 1997 g. № 188] rezhim dostupa: http://base.consultant.ru/cons/cgi/online.cgi?req=doc;base=LAW;n=182734 (data obrashhenija: 25.07.2015).

7. O personal'nyh dannyh: feder. zakon FZ-152 [podp. Prezidentom Rossijskoj Federacii 27 ijulja 2006 g.] rezhim dostupa: http://base.garant.ru/10200083/#friends (data obrashhenija: 25.07.2015).