ВЕСТНИК НОВЫХ МЕДИЦИНСКИХ ТЕХНОЛОГИЙ – 2016 – N 1
Электронный журнал

УДК: 303.722.8; 519.683.5; 614.1 DOI: 10.12737/18445

О РЕАЛИЗАЦИИ АЛГОРИТМА РАСПОЗНАВАНИЯ ТЕКСТА В БАЗЕ ДАННЫХ

В.А. ХРОМУШИН[*], О.В. ХРОМУШИН[**], К.Ю. КИТАНИНА[*]

[*]*Тульский государственный университет, ул. Болдина, д. 128, Тула, Россия, 300028, e-mail: vik@khromushin.com*
[**]*Тульское отделение Академии медико-технических наук, а/я 1842, Тула, Россия, 300026*

Аннотация. В статье рассматриваются особенности реализации алгоритма распознавания текста методом «скользящего увеличивающегося окна», используемого для кодирования множественных причин смерти. Используемый алгоритм динамически «настраивает» степень совпадения и находит наиболее похожий вариант, а также позволяет распознавать текст с грамматическими ошибками и с переставленными словами в формулировке причины смерти.

В статье представлены три варианта реализации алгоритма распознавания текста, увеличивающих быстродействие. Первый вариант основан на исключении одного из цикла путем замены его одновременным вычислением с различными размерами окна (от 1 до 16). Второй вариант основан на предварительной фильтрации, например, сканированием по трем буквам, и использование промежуточной базы для размещения в ней фильтрованной информации. Этот вариант позволяет уменьшить объем сортируемой информации и за счет этого увеличивает быстродействие. Третий вариант также основан на фильтрации и заключается в сортировке информации в запросе, выполненном на базе предыдущего запроса с фильтрацией информации. По каждому варианту реализации указаны достоинства и недостатки. Оценка результата оценивалась по быстродействию и правильности распознавания. При этом база насчитывала 8472 формулировок, предназначенных для кодирования множественных причин смерти.

Изложенный анализ путей реализации полезен в разработке программного модуля, используемого в регистре смертности населения. Рекомендуется третий вариант, основанный на фильтрации, для реализации на языке *Visual C++*.

Ключевые слова: распознавание, алгоритм, база данных, поиск, оценка.

ABOUT THE USE OF THE RECOGNITION ALGORITHM OF THE TEXT IN DATABASE

V.A. KHROMUSHIN[*], O.V. KHROMUSHIN[**], K.YU. KITANINA[*]

[*]*Tula State University, 128, Boldin str., Tula, Russia, 300028, e-mail: vik@khromushin.com*
[**]*Tula Branch of the Academy of Medical and Technical Sciences, PO Box 1842, Tula, Russia, 300026*

Abstract. This article presents the features of the use of the recognition algorithm of the text by method "slitherring widening window" for coding the plural reasons to deaths. The used algorithm dynamically "adjusts" degree of the coincidence and finds the most similar variant, as well as allows to recognize the text with grammatical errors and with ceased word in wording of the reason to deaths.

The authors propose three variants to realization of the recognition algorithm of the text, that increase the speed of action. The first variant is based on the elimination of one of its replacement cycle by calculating the simultaneous windows with different sizes (1 to 16). The second variant is based on the pre-filter, for example, by scanning three letters, and the use of the intermediate base for receiving the filtered information. This option allows you to reduce the amount of sorted data and thereby increases performance. The third variant based on the filtering, involves sorting information in the query executed on the basis of previous information query filtration. The advantages and disadvantages identified for each option. The results were evaluated on the speed of action and accuracy of recognition. In this framework, there were 8472 languages used for encoding of multiple causes of death.

The described analysis is useful in developing a software module used to register mortality. It is recommended the third option, based on the filter, to implement the language *Visual C ++*.

Key words: recognition, algorithm, database, searching, estimation.

Введение. Мониторинг смертности является важным звеном в оценке здоровья населения. Информация о смертности позволяет органам общественного здравоохранения правильно фокусировать свою деятельность. Принципы реализации мониторинга смертности предусматривают наряду с обеспечением точности кодирования множественных причин смерти средства распознавания текста формулировки причины смерти, что облегчает противопоставление этой формулировке кода

международной статистической классификации болезней и проблем, связанных со здоровьем (МКБ-Х) [1, 10-12, 14].

В здравоохранении Тульской области режим распознавания текста используется в регистре смертности, наряду с автоматическим определением первоначальной причины смерти (аналогично программному обеспечению США) [1, 2, 9-14, 17]. В этом программном обеспечении по вводимой формулировке осуществляется поиск наиболее похожего текста, который вместе с кодом заносится в медицинское свидетельство о смерти. Аналогичный режим используется в пакете программ, используемом в США (разработчик – фирма *CDC*). Правильное кодирование является основой обеспечения достоверности информации о смертности населения и выполнения аналитических исследований [3-8, 15, 16, 19].

В основе алгоритма распознавания текста положен метод «скользящего увеличивающегося окна». Используемый алгоритм динамически «настраивает» степень совпадения и находит наиболее похожий вариант, а также позволяет распознавать текст с грамматическими ошибками и с переставленными словами в формулировке причины смерти [9-12].

Объекты и методы исследования. Объектом данного исследования является алгоритм распознавания, основанный на методе «скользящего увеличивающегося окна». Для облегчения восприятия исходного алгоритма представим его в виде следующей функции в среде Access [18]:

```
Function RpzText(KodB As Variant, SprB As Variant) As Double
'SprB Строковое выражение, в котором производится поиск
'KodB Искомое строковое выражение
  Dim i As Long, u As Long, XR As Double, pos As Long, TempStr As String
  XR = 0
  For u = 1 To Len(KodB)  'величина окна u
    For i = 1 To Len(KodB) - u + 1  'последовательное сканирование
      TempStr = Mid(KodB, i, u)  'выделение фрагмента в окне шириной u
      pos = InStr(1, SprB, TempStr, vbTextCompare)  'поиск
      If pos > 0 And Len(KodB) >= (i + u - 1) Then XR = XR + u ^ k,  'накопление результирующей
весовой оценки
    Next i
  Next u
  RpzText = XR / Len(SprB)
End Function
'где k рекомендуется брать равным 2,6
```

В регистре смертности эта процедура реализована на языке *Visual C++*.

Данная процедура в регистре смертности вызывается двойным кликом по вводимой формулировке тогда, когда пользователь начинает не видеть в подключенном к полю списке текст формулировки. Учитывая, что база синонимов достаточно большая, находить код вручную требует значительных затрат времени.

В результате распознавания пользователю выдаются наиболее вероятные варианты, представленные в порядке убывания по степени похожести [18].

Оценка результата оценивалась по быстродействию и правильности распознавания. При этом база насчитывала 8472 формулировок с кодами МКБ-Х, предназначенных для кодирования множественных причин смерти. В качестве тестовой формулировки было использовано выражение в двойной ошибкой: «Сахарный диабет лобильный» (с переставленными словами и грамматической ошибкой), в то время как искомая формулировка была: «Диабет сахарный лабильный».

Исходный алгоритм является оптимизированным и подробно изложен в статье [18]. Реализация исходного алгоритма по схеме «Цикл в цикле» ухудшает быстродействие.

Обсуждение результатов. Реализация исходного алгоритма с высоким быстродействием возможна тремя способами.

1. Первый способ основан на исключении одного из цикла путем замены его одновременным вычислением с различными размерами окна (от 1 до 16):

```
Function RText(KodB As Variant, SprB As Variant) As Double
  Dim i As Long, u As Long, XR As Double, TempStr As String, pos As Long
  XR = 0
    For i = 1 To Len(KodB)
      TempStr = Mid(KodB, i, 1)
      pos = InStr(1, SprB, TempStr, vbTextCompare)
      If pos > 0 And Len(KodB) >= i Then XR = XR
      TempStr = Mid(KodB, i, 2)
      pos = InStr(1, SprB, TempStr, vbTextCompare)
```

```
        If pos > 0 And Len(KodB) >= (i + 1) Then XR = XR + (2 ^ 2.6)
                                    ...
        TempStr = Mid(KodB, i, 12)
        pos = InStr(1, SprB, TempStr, vbTextCompare)
        If pos > 0 And Len(KodB) >= (i + 15) Then XR = XR + (16 ^ 2.6)
      Next i
    RText = XR / Len(SprB)
  End Function
```

Такой прием, несмотря на ограничения на максимальный размер окна, практически не сказывается на эффективности распознавания. В тоже время он дает повышение быстродействия примерно на 17% на выбранном тестовом примере.

В представленном варианте реализации алгоритма распознавания «узким» местом в быстродействии является сортировка по убыванию при отображении результата. Если уменьшить объем сортируемой информации, то быстродействие должно увеличиться.

2. Второй способ основан на предварительной фильтрации, например, по трем буквам. В этом случае мы уменьшаем объем сортируемой информации и, следовательно, можем рассчитывать на увеличение быстродействия.

Реализация этого способа может быть осуществлена с помощью следующей фильтрующей функции:

```
Function FRText(KodB As Variant, SprB As Variant) As Long
  Dim i As Long, TempStr As String, pos As Long
  FRText = 0
    For i = 1 To Len(KodB) - 3
      TempStr = Mid(KodB, i, 3)
      pos = InStr(1, SprB, TempStr, vbTextCompare)
      If pos > 0 Then FRText = 1
    Next i
End Function
```

Важно отметить, что фильтрацию по трем буквам в данном варианте осуществляется сканирование окном размером в три знакоместа. Здесь нельзя использовать более простую и быструю фильтрацию по первым трем буквам, поскольку пользователь может переставить слова и даже допустить грамматическую ошибку в первом слоге.

Поскольку фильтрация осуществляется по трем буквам, то из 16 вычислений можно исключить сканирование окном в три знакоместа.

Реализовать этот способ можно добавлением отфильтрованной информации в промежуточную базу. Перед добавлением необходимо из нее удалить все данные от предыдущего цикла работы. Сортировка по убыванию для отображения результата будет осуществляться из этой промежуточной базы, что средствами *Access* производится достаточно быстро. «Узким» местом в обеспечении быстродействия в этом варианте является процесс добавления отфильтрованной информации в промежуточную базу.

3. Третий способ также основан на фильтрации и заключается в сортировке информации в запросе выполненном на базе предыдущего запроса с фильтрацией информации. В этом способе за счет уменьшенного числа записей после фильтрации сортировка осуществляется быстрее.

Необходимо отметить, что способы 2 и 3 примерно равноценны по быстродействию. Небольшое преимущество в данном случае имеет третий способ.

Реализация последнего способа за языке *Visual C++* позволило обеспечить высокое быстродействие. В результате пользователь не замечает задержки вывода результата распознавания.

Выводы:

1. Изложенный анализ путей реализации полезен в разработке программного модуля, используемого в регистре смертности населения.
2. Рекомендуется третий способ, основанный на фильтрации, для реализации на языке *Visual C++*.

Литература

1. Вайсман Д.Ш., Погорелова Э.И., Хромушин В.А. О создании автоматизированной комплексной системы сбора, обработки и анализа информации о рождаемости и смертности в Тульской области // Вестник новых медицинских технологий. 2001. №4. С. 80–81.

2. Вайсман Д.Ш., Никитин С.В., Погорелова Э.И., Секриеру Е.М., Хромушин В.А. Повышение достоверности кодирования внешних причин смерти // Вестник новых медицинских технологий. 2006. Т.13, №1. С. 147–148.

3. Даильнев В.И., Хромушин В.А., Китанина К.Ю. Анализ смертности населения Тульской области от болезней системы кровообращения // Вестник новых медицинских технологий (электронное издание). 2013. №1. Публикация 2-15. URL: http://medtsu.tula.ru/VNMT/Bulletin/E2013-1/4210.pdf (дата обращения 10.01.2013).

4. Макишева Р.Т., Хадарцев А.А., Хромушин В.А., Даильнев В.И. Возрастной анализ смертности населения Тульской области от сахарного диабета // Вестник новых медицинских технологий (электронное издание). 2014. №1. Публикация 7-9. URL: http://medtsu.tula.ru/VNMT/Bulletin/E2014-1/4900.pdf. (дата обращения 06.08.2014). DOI:10.12737/5613.

5. Макишева Р.Т., Хромушин В.А., Прилепа С.А., Ластовецкий А.Г. Гендерные особенности смертности больных сахарным диабетом в Тульской области // Вестник новых медицинских технологий. 2015. Т. 22, №2. С. 60–67. DOI:10.12737/11835.

6. Погорелова Э.И., Секриеру Е.М., Стародубов В.И., Мелехина Л.Е., Нотсон Ф.К., Хромушин В.А., Вайсман Д.Ш., Мельников В.А., Дегтерева М.И., Одинцова И.А., Корчагин Е.Е., Виноградов К.А. Заключительный научный доклад «Разработка системы мероприятий для совершенствования использования статистических данных о смертности населения Российской Федерации» (Международный исследовательский проект 1АХ202)». Москва: ЦНИИ организации и информатизации МЗ РФ, 2003. 34 с.

7. Китанина К.Ю., Хромушин В.А. Анализ инвалидности населения Тульской области // Вестник новых медицинских технологий (электронное издание). 2012. №1. URL: http://medtsu.tula.ru/VNMT/Bulletin/E2012-1/3717.pdf (дата обращения 19.01.2012).

8. Хадарцев А.А., Хромушин В.А., Андреева Ю.В., Даильнев В.И. Анализ смертности от сахарного диабета 2 типа в Тульской области // Вестник новых медицинских технологий. 2012. Т.XIX, №3. С. 164–167.

9. Хромушин В.А., Китанина К.Ю., Даильнев В.И. Кодирование множественных причин смерти // Учебное пособие. Тула: Изд-во ТулГУ, 2012. 60 с.

10. Хромушин В.А., Хадарцев А.А., Даильнев В.И., Ластовецкий А.Г. Принципы реализации мониторинга смертности на региональном уровне // Вестник новых медицинских технологий (электронное издание). 2014. №1. Публикация 7-6. URL: http://medtsu.tula.ru/VNMT/Bulletin/E2014-1/4897.pdf (дата обращения 06.08.2014). DOI:10.12737/5610.

11. Хромушин В.А. Системный анализ и обработка информации медицинских регистров в регионах // Автореферат диссертации на соискание ученой степени доктора биологических наук. Тула: Научно-исследовательский институт новых медицинских технологий, 2006. 44 с.

12. Хромушин В.А., Черешнев А.В., Честнова Т.В. Информатизация здравоохранения. Учебное пособие. Тула: Изд-во ТулГУ, 2007. 207 с.

13. Хромушин В.А., Вайсман Д.Ш. Мониторинг смертности с международной сопоставимостью данных // В сборнике тезисов докладов научно-практической конференции "Современные инфокоммуникационные технологии в системе охраны здоровья". 2003 Нояб. 13-14., Москва. С. 122.

14. Хромушин В.А. Методология обработки информации медицинских регистров. Монография. Тула: Изд-во ТулГУ, 2004. 120 с.

15. Хромушин В.А. Методология анализа множественных причин смерти // Вестник новых медицинских технологий. 2004. №3. С. 107–109.

16. Хромушин В.А., Хадарцев А.А., Даильнев В.И., Китанина К.Ю. Анализ динамики смертности возрастных когорт населения Тульской области // Вестник новых медицинских технологий. 2014. №1. Публикация 7-5. URL: http://medtsu.tula.ru/VNMT/Bulletin/E2014-1/4896.pdf (дата обращения 06.08.2014). DOI:10.12737/5609.

17. Хромушин В.А., Погорелова Э.И., Секриеру Е.М. Возможности дополнительного повышения достоверности данных по смертности населения // Вестник новых медицинских технологий. 2005. Т.12, №2. С. 95–96.

18. Хромушин В.А. Анализ алгоритма распознавания текста в базе данных // Вестник новых медицинских технологий. 2013. Т. 20, №3. С. 13–16.

19. Щеглов В.Н., Бучель В.Ф., Хромушин В.А. Логические модели структур заболеваний за 1986-1999 годы участников ликвидации аварии на ЧАЭС и/или мужчин, проживающих в пораженной зоне и имеющих злокачественные новообразования органов дыхания // Радиация и риск. 2002. №13. С. 57–59.

References

1. Vaysman DS, Pogorelova EI, Khromushin VA. O sozdanii avtomatizirovannoy kompleksnoy sistemy sbora, obrabotki i analiza informatsii o rozhdaemosti i smertnosti v Tul'skoy oblasti. Vestnik novykh meditsinskikh tekhnologiy. 2001;4:80-1. Russian.

2. Vaysman DS, Nikitin SV, Pogorelova EI, Sekrieru EM, Khromushin VA. Povyshenie dostovernosti kodirovaniya vneshnikh prichin smerti. Vestnik novykh meditsinskikh tekhnologiy. 2006;8(1):147-8. Russian.

3. Dail'nev VI, Khromushin VA, Kitanina KY. Analiz smertnosti naseleniya Tul'skoy oblasti ot bolezney sistemy krovoobrashcheniya. Vestnik novykh meditsinskikh tekhnologiy (Elektronnoe izdanie). 2013 [cited 2013 Jan 10];1: [about 5 p.]. Russian. Available from: http://medtsu.tula.ru/VNMT/Bulletin/E2013-1/4210.pdf.

4. Makisheva RT, Khadartsev AA, Khromushin VA, Dail'nev VI. Vozrastnoy analiz smertnosti naseleniya Tul'skoy oblasti ot sakharnogo diabeta. Vestnik novykh meditsinskikh tekhnologiy (Elektronnoe izdanie). 2014 [cited 2014 Aug 06];1: [about 11 p.]. Russian. Available from: http://medtsu.tula.ru/VNMT/Bulletin/E2014-1/4900.pdf. DOI:10.12737/5613.

5. Makisheva RT, Khromushin VA, Prilepa SA, Lastovetskiy AG. Gendernye osobennosti smertnosti bol'nykh sakharnym diabetom v Tul'skoy oblasti. Vestnik novykh meditsinskikh tekhnologiy. 2015;22(2):60-7. DOI:10.12737/11835. Russian.

6. Pogorelova EI, Sekrieru EM, Starodubov VI, Melekhina LE, Notson FK, Khromushin VA, Vaysman DS, Mel'nikov VA, Degtereva MI, Odintsova IA, Korchagin EE, Vinogradov KA. Zaklyuchitel'nyy nauchnyy doklad «Razrabotka sistemy meropriyatiy dlya sovershenstvovaniya ispol'zovaniya statisticheskikh dannykh o smertnosti naseleniya Rossiyskoy Federatsii» (Mezhdunarodnyy issledovatel'skiy proekt 1AKh202)». Moscow: TsNII organizatsii i informatizatsii MZ RF, 2003. Russian.

7. Kitanina KY, Khromushin VA. Analiz invalidnosti naseleniya Tul'skoy oblasti. Vestnik novykh meditsinskikh tekhnologiy (Elektronnoe izdanie). 2012 [cited 2012 Jan 19];1: [about 15 p.]. Russian. Available from: http://medtsu.tula.ru/VNMT/Bulletin/E2012-1/3717.pdf.

8. Khadartsev AA, Khromushin VA, Andreeva YV, Dail'nev VI. Analiz smertnosti ot sakharnogo diabeta 2 tipa v Tul'skoy oblasti. Vestnik novykh meditsinskikh tekhnologiy. 2012; 19(3):164-7.

9. Khromushin VA, Kitanina KY, Dail'nev VI. Kodirovanie mnozhestvennykh prichin smerti. Uchebnoe posobie. Tula: Izd-vo TulGU, 2012. Russian.

10. Khromushin VA, Khadartsev AA, Dail'nev VI, Lastovetskiy AG. Printsipy realizatsii monitoringa smertnosti na regional'nom urovne. Vestnik novykh meditsinskikh tekhnologiy (Elektronnoe izdanie). 2014 [cited 2014 Aug 06];1 [about 7 p.]. Russian. Available from: http://medtsu.tula.ru/VNMT/Bulletin/E2014-1/4897.pdf. DOI:10.12737/5610.

11. Khromushin VA. Sistemnyy analiz i obrabotka informatsii meditsinskikh registrov v regionakh [dissertation]. Tula (Tula region): Nauchno-issledovatel'skiy institut novykh meditsinskikh tekhnologiy; 2006. Russian.

12. Khromushin VA, Chereshnev AV, Chestnova TV. Informatizatsiya zdravookhraneniya. Uchebnoe posobie. Tula: Izd-vo TulGU, 2007. Russian.

13. Khromushin VA, Vaysman DS. Monitoring smertnosti s mezhdunarodnoy sopostavimost'yu dannykh. V sbornike tezisov dokladov nauchno-prakticheskoy konferentsii «Sovremennye infokommunikatsionnye tekhnologii v sisteme okhrany zdorov'ya»; 2003 Nov.13-14; Moscow, p. 122. Russian.

14. Khromushin VA. Metodologiya obrabotki informatsii meditsinskikh registrov. Monografiya. Tula: Izd-vo TulGU, 2004. Russian.

15. Khromushin VA. Metodologiya analiza mnozhestvennykh prichin smerti. Vestnik novykh meditsinskikh tekhnologiy. 2004;3:107-9. Russian.

16. Khromushin VA, Khadartsev AA, Dail'nev VI, Kitanina KY. Analiz dinamiki smertnosti vozrastnykh kogort naseleniya Tul'skoy oblasti. Vestnik novykh meditsinskikh tekhnologiy. 2014 [cited 2014 Aug 06];1 [about 14 p.] Russian. Available from: http://medtsu.tula.ru/VNMT/Bulletin/E2014-1/4896.pdf. DOI:10.12737/5609.

17. Khromushin VA, Pogorelova EI, Sekrieru EM. Vozmozhnosti dopolnitel'nogo povysheniya dostovernosti dannykh po smertnosti naseleniya. Vestnik novykh meditsinskikh tekhnologiy. 2005; 8(2):95-6. Russian.

18. Khromushin VA. Analiz algoritma raspoznavaniya teksta v baze dannykh. Vestnik novykh meditsinskikh tekhnologiy. 2013;20(3):13-6. Russian.

19. Shcheglov VN, Buchel' VF, Khromushin VA. Logicheskie modeli struktur zabolevaniy za 1986-1999 gody uchastnikov likvidatsii avarii na ChAES i/ili muzhchin, prozhivayushchikh v porazhennoy zone i imeyushchikh zlokachestvennye novoobrazovaniya organov dykhaniya. Radiatsiya i risk. 2002;13:57-9. Russian.

Библиографическая ссылка:
Хромушин В.А., Хромушин О.В., Китанина К.Ю. О реализации алгоритма распознавания текста в базе данных // Вестник новых медицинских технологий. Электронное издание. 2016. №1. Публикация 1-1. URL: http://www.medtsu.tula.ru/VNMT/Bulletin/E2016-1/1-1.pdf (дата обращения: 29.02.2016). DOI: 10.12737/18445.