

Институт программных систем имени А.К. Айламазяна РАН

Способ **Проблема** повышения качества алгоритмов,
~~базирующихся на метрике Левенштейна или~~
~~родственных понятиях~~
выделения различий в текстовых файлах

С.В. Знаменский

E-mail: svz@latex.pereslavl.ru

Прикладная сторона проблемы

Выделение различий для анализа

Выделение различий для автоматической обработки

Математические модели

Неадекватность LCS

Неадекватность patiencediff

Неадекватность Bdiff

Проблема:

Найти алгоритм, лаконично, адекватно и за приемлемое время выделяющий различия между двумя версиями текстового файла.

Выделение различий между двумя версиями используется вручную множеством программных продуктов для просмотра и редактирования

Beyond Compare	Scotter Software ^[4]	Proprietary	\$US30 (std), 50 (pro)	1996	2014 (v4.0)
Compare++	Coode Software ^[2]	Proprietary	\$29.95, 30-day free trial ^[3]	2010	2013 (v2.15)
diff, diff3	AT&T	BSD 3-clause, BSD 4-clause, CDDL, GPL, Proprietary	-	1974	
Ediff	Michael Kifer ^[5]	GPL	Free	1994	2.81.4
ExamDiff Pro	PrestoSoft ^[6]	Proprietary	\$35, free file-only version ^[7]	1998	2014 (v7.0.1.6)
Far Manager (compare)	Eugene Roshal (original); FAR Group	Revised BSD license	Free	1996	August 2014 (v 3.0 build 4040)
fc	Microsoft ^[8]	Proprietary	Part of OS	1987	
FileMerge	Apple inc.	Proprietary	Free (part of Apple Developer Tools)	1993 (part of NEXTSTEP 3.2 ^[9])	2014 (v2.8)
Guffy SureMerge	Guffy Software ^[10]	Proprietary	\$37.50(Pro), 75(expert)	2000	July, 2014 (v10.6)
IntelliJ IDEA (compare)	JetBrains ^[11]	Proprietary	Part of application		2007 (v6.0.5)
jEdit JDiff plugin	Various ^[12]	GPL	Free	1998	2012 (3.3.0)
Kompare	Otto Bruggeman ^[13]	GPLv2+	Free		
Lazarus Diff	Lazarus (software)	GPL	-	2000	2009
Meld	Stephen Kennedy ^[15]	GPLv2+	-	2002	2014-1-23 (v1.8.4)
Perforce P4Merge	Perforce	Proprietary ^[17]	-		2013 (2013.2)
Pretty Diff	Austin Cheney ^[18]	GPL-compatible	Free	2009	2012-07-05
Tkdiff	Tkdiff ^[19]	GPLv2+	Free	2003 (or before)	2011-11-27 (v4.2)
Total Commander (compare)	Christian Ghisler ^[20]	Proprietary	Part of application		2010-12-17 (v7.56a)
twdiff (TextWrangler Diff Helper)	Bare Bones Software, inc. ^[21]	Proprietary	free (with TextWrangler [free])		2012 (1.0 (v22))
vimdiff	Bram Moolenaar et al.	GPL-compatible	-	2001	2006-05-07 (v7.0)
WinDiff	Microsoft ^[22]	Proprietary	Part of Platform SDK	1992	2003 ? (v6.1)
WinMerge	Dean Grimm ^[23]	GPL	Free	1998	2013-02-03 (v2.14.0)

для отслеживания изменений в исходных файлах

ЗАЧЕМ и КАКИЕ нужны суперкомпьютеры эксафлопсного ~~эксафлопсного~~ класса?

Предсказательное моделирование свойств и многомасштабных процессов в материаловедении

Аннотация. Рассматривается подход, позволяющий выявить, для каких задач нужны суперкомпьютеры эксафлопсного класса. Возможности подхода рассмотрены на примерах актуальных задач материаловедения, физики конденсированного вещества и плотной плазмы, для решения которых необходимо атомистическое моделирование на современных и создаваемых в настоящее время суперкомпьютерах. Для каждой задачи проведено соответствие между набором изучаемых явлений и требуемым уровнем быстродействия (числа ядер) вычислительной системы. Показана масштабируемость параллельных программ моделирования и перспектива расширения предсказательной способности методов по мере увеличения числа вычислительных ядер и/или использования специализированных архитектур (графические ускорители). Рассмотрена иерархия методов моделирования, необходимых для адекватного описания свойств веществ на различных пространственных и временных масштабах. На наиболее глубоком нанометровом/пикометровом масштабе для моделирования электронной динамики и построения эффективных потенциалов взаимодействия частиц применяется теория функционала плотности (квантовая молекулярная динамика). Классический метод молекулярной динамики позволяет явно рассмотреть системы движущихся атомов вплоть до ~~микро-масштабов~~ ~~микромасштабов~~. Выход на ~~макно-масштабы~~ ~~макромасштабы~~ осуществляется с помо-

или в иных текстах при совместной подготовке

Kompare

File Difference Settings Help

Source Folder	Destination Folder	Source File	Destination File	Source Line	Destination Line	Difference
/home/robamler/Documents/	/home/robamler/Documents/	hitchhiker-040624	hitchhiker-current	32	48	Changed 1 line
				38	56	Changed 1 line
				51	69	Changed 1 line
				57	75	Inserted 2 lines
				59	79	Deleted 1 line

hitchhiker-040624	hitchhiker-current
52	66 * [[iso long, and thanks for all the fish]]' ([[1991]])
53 [[Neil Gaiman]] has written ''[[Don't Panic: The Official]]	67 * ''[[Mostly Harmless]]'' ([[1992]])
54	68
55 A novel, ''[[Douglas Adams's Starship Titanic]]'', based	69 A short story was also written, ''[[Young Zaphod Plays I]]
56	70
57 Two collaborative [[Internet]] projects were inspired by	71 [[Neil Gaiman]] has written ''[[Don't Panic: The Official]]
58	72
59 ==Characters==	73 A novel, ''[[Douglas Adams's Starship Titanic]]'', based
60	74
61 * [[Arthur Dent]]	75 Nowbagger the Infinitely Prolonged, a character from ''[[
62 * [[Ford Prefect (HHG) Ford Prefect]]	76
63 * [[Trillian]]	77 Two collaborative [[Internet]] projects were inspired by
64 * [[Zaphod Beeblebrox]]	78
65 * [[Marvin the Paranoid Android]]	79
66 * [[Vogon]]s	80 [[HitchhikerBooks]]
67 * [[Deep Thought]]	81
68 * [[Concepts and minor characters from The Hitchhiker's	82
69	83 ==See also==
70 ==See also==	84
71	85 * [[Dirk Gently]] -- another [[Douglas Adams]] creation
72 * [[Places in The Hitchhikers Guide to the Galaxy Places	86 * [[Hitchhiking]]
73 * [[Infinite Improbability Drive]]	87 * [[H2G2 h2g2]]
74 * [[Total Perspective Vortex]]	88
75 * [[The Answer to Life, the Universe, and Everything]]	89 ==External links==
76 * [[Dirk Gently]] -- another [[Douglas Adams]] creation	90
77 * [[SEP field]]	91 * http://www.bbc.co.uk/cult/hitchhikers/ ''The Hitchhik
78 * [[Hitchhiking]]	92 * http://www.hitchhikersmovie.com ''The Hitchhiker's G
79 * [[Sirius Cybernetics Corporation]]	93 * The http://www.cinescape.com/0/editorial.asp?obj_id=3
80 * [[H2G2]]	94 * http://www.imdb.com/title/tt0371724/ The film's status
81 * [[42 (number) 42]]	95 * http://www.abovethetitle.com/content/programmes/hitchhiker
82	96

Comparing file file:/home/robamler/Documents/hitchhiker-...th file file:/home/robamler/Documents/hitchhiker-current | 1 of 23 differences, 0 applied | 1 of 1 file

— как построчно,

Some widely used file comparison programs are `diff`, `cmp` (Unix), `FileMerge` (Apple Developer Tools), `WinMerge`, `Beyond Compare`, and `Microsoft File Compare`.

Line 8:

== Method types ==

Most file comparison tools find the `longest common subsequence` between two files. Any data not in the longest common subsequence is presented as an insertion or deletion.

In 1978, Paul Heckel published an algorithm that identifies most moved blocks of text.
 {{Citation|last1=Viégas|first1=Paul|last1=Heckel|year=1978|title=A Technique for Isolating Differences Between Files|journal=Communications of the ACM|volume=21|pages=264–268|url=http://documents.scribd.com/docs/10ro9oowpo1h81pgh1as.pdf|accessdate=2011-12-04|doi=10.1145/359460.359467}} This is

Some widely used file comparison programs are `diff`, `cmp` (Unix), `FileMerge` (Apple Developer Tools), `WinMerge`, `Beyond Compare`, and `Microsoft File Compare`.

Line 8:

== Method types ==

Most file comparison tools find the **longest common problem** `longest common subsequence` between two files. Any data not in the longest common subsequence is presented as an insertion or deletion.

In 1978, Paul Heckel published an algorithm that identifies most moved blocks of text.
 {{Citation|last1=Viégas|first1=Paul|last1=Heckel|year=1978|title=A Technique for Isolating Differences Between Files|journal=Communications of the ACM|volume=21|pages=264–268|url=http://documents.scribd.com/docs/10ro9oowpo1h81pgh1as.pdf|accessdate=2011-12-04|doi=10.1145/359460.359467}} This is

так и более детально (википедия)


```

psta-view.log (~/.archive/our_pub/sit.2014/2014/14-2-Paimushkin) (1 of 2) - GVIMS
<use pic3.pdf>
Package pdfTeX.def Info: pic3.pdf used on input line 208.
(pdfTeX.def) Requested size: 315.82558pt x 240.0351
Package babel Info: Redefining russian shorthand |
(babel) in language append on input line 212.
+... 8 lines: Package babel Info: Redefining russian shorthand
Package babel Info: Redefining russian shorthand ~
(babel) in language append on input line 221.
[143 <./pic2.pdf> <./pic3.pdf

pdfTeX warning: pdflatex (file ./pic3.pdf): PDF inclusion: multi
age group included in a single page
>] <pic5.pdf, id=118, 290.5023pt x 128.59946pt>
File: pic5.pdf Graphic file (type pdf)
<use pic5.pdf>
Package pdfTeX.def Info: pic5.pdf used on input line 223.
(pdfTeX.def) Requested size: 315.82558pt x 139.8095
)
Runaway argument?
+... 34 lines: (http://ru.wikipedia.org/wiki/\section (http://ru
Package babel Info: Redefining russian shorthand |
(babel) in language append on input line 9.
Package babel Info: Redefining russian shorthand ~
(babel) in language append on input line 9.
[145]
Package atveryend Info: Empty hook `AfterLastShipout' on input
(./psta-view.aux)
Package atveryend Info: Executing hook `AtVeryEndDocument' on i
Package atveryend Info: Executing hook `AtEndAfterFileList' on
Package rerunfilecheck Info: File `psta-view.out' has not chan
(rerunfilecheck) Checksum: 5903E3BBD1EF2A614480B05
Package atveryend Info: Empty hook `AtVeryVeryEnd' on input lin
)
Here is how much of TeX's memory you used:
13938 strings out of 493304
189784 string characters out of 6139871
273749 words of memory out of 5000000
16928 multiletter control sequences out of 15000+600000
30853 words of font info for 68 fonts, out of 8000000 for 9000
1119 hyphenation exceptions out of 8191
461.7n,40p,354h,551s stack positions out of 5000i,500n,10000p,
/usr/share/texmf/fonts/enc/dvips/cn-super/cn-super-12a.enc)<u
psta-view.log,unix-latn1.ru, 50 43019(1,973=89%(1085))

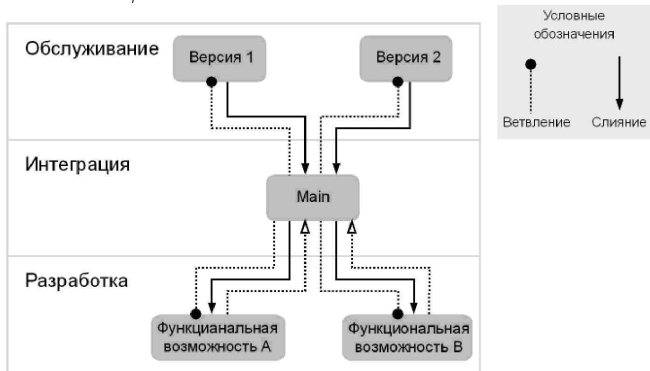
<use pic3.pdf>
Package pdfTeX.def Info: pic3.pdf used on input line 208.
(pdfTeX.def) Requested size: 315.82558pt x 240.03
Package babel Info: Redefining russian shorthand |
(babel) in language append on input line 212.
+... 8 lines: Package babel Info: Redefining russian shorthan
Package babel Info: Redefining russian shorthand ~
(babel) in language append on input line 221.
[143 <./pic2.pdf> <./pic3.pdf

pdfTeX warning: pdflatex (file ./pic3.pdf): PDF inclusion: mu
age group included in a single page
>] <pic5.pdf, id=118, 290.5023pt x 128.59946pt>
File: pic5.pdf Graphic file (type pdf)
<use pic5.pdf>
Package pdfTeX.def Info: pic5.pdf used on input line 223.
(pdfTeX.def) Requested size: 315.82558pt x 139.80
)
Runaway argument?
+... 34 lines: (http://ru.wikipedia.org/wiki/\section (http://
Package babel Info: Redefining russian shorthand |
(babel) in language append on input line 9.
Package babel Info: Redefining russian shorthand ~
(babel) in language append on input line 9.
[145]
Package atveryend Info: Empty hook `AfterLastShipout' on inpu
(./psta-join.aux)
Package atveryend Info: Executing hook `AtVeryEndDocument' on
Package atveryend Info: Executing hook `AtEndAfterFileList' o
Package rerunfilecheck Info: File `psta-join.out' has not cha
(rerunfilecheck) Checksum: 5903E3BBD1EF2A614480B05
Package atveryend Info: Empty hook `AtVeryVeryEnd' on input l
)
Here is how much of TeX's memory you used:
13938 strings out of 493304
189784 string characters out of 6139871
273749 words of memory out of 5000000
16928 multiletter control sequences out of 15000+600000
30853 words of font info for 68 fonts, out of 8000000 for 90
1119 hyphenation exceptions out of 8191
461.7n,40p,354h,551s stack positions out of 5000i,500n,10000
pdfTeX warning (dest): name{russiancontents} has been referre
psta-join.log,unix-latn1.ru, 50 43019(1,973=89%(1091))

```

для отслеживания изменений в логах

Используется системами сборки сложного программного обеспечения для выявления конфликтов и наложения изменений,



независимо внесённых разными разработчиками. Качество и сложность сборки сильно зависит от точности локализации изменений.

Основные математические модели:

Широко используются (по Baudiš P. Current concepts in version control systems //arXiv preprint arXiv:1405.3496. – 2014.(2009-09-11)):

1. Longest Common Subsequence LCS (Myers E.: An $O(ND)$ Difference Algorithm and its Variations, *Algorithmica* 1 (1986) 251–266) — GNU diffutils, perl, ...
2. patiencediff (LCS for lines which occur exactly once on both sides, then recurse between). Cohen B., et al.: Bazaar Source Code, bzrlib/patiencediff.py, <http://www.bazaar-vcs.org/Download>
3. Bdiff longest common continuous substring — Ratcliff J. W., Metzener D. E. Pattern-matching-the gestalt approach //Dr Dobbs Journal. – 1988. – Т. 13. – №. 7. – С. 46-diffliб(python), Mercurial, Revlog

не вполне адекватны

Основные математические модели:

Широко используются (по Baudiš P. Current concepts in version control systems //arXiv preprint arXiv:1405.3496. – 2014.(2009-09-11)):

1. Longest Common Subsequence LCS (Myers E.: An $O(ND)$ Difference Algorithm and its Variations, *Algorithmica* 1 (1986) 251–266) — GNU diffutils, perl, ...
2. *patiencediff* (LCS for lines which occur exactly once on both sides, then recurse between). Cohen B., et al.: Bazaar Source Code, bzrlib/patiencediff.py, <http://www.bazaar-vcs.org/Download>
3. *Bdiff* longest common continuous substring — Ratcliff J. W., Metzener D. E. Pattern-matching-the gestalt approach //Dr Dobbs Journal. – 1988. – Т. 13. – №. 7. – С. 46-
difflib(python), Mercurial, Revlog

не вполне адекватны

Основные математические модели:

Широко используются (по Baudiš P. Current concepts in version control systems //arXiv preprint arXiv:1405.3496. – 2014.(2009-09-11)):

1. Longest Common Subsequence LCS (Myers E.: An $O(ND)$ Difference Algorithm and its Variations, *Algorithmica* 1 (1986) 251–266) — GNU diffutils, perl, ...
2. *patience*diff (LCS for lines which occur exactly once on both sides, then recurse between). Cohen B., et al.: Bazaar Source Code, bzrlib/*patience*diff.py, <http://www.bazaar-vcs.org/Download>
3. Bdiff longest common continuous substring — Ratcliff J. W., Metzener D. E. Pattern-matching-the gestalt approach //Dr Dobbs Journal. – 1988. – Т. 13. – №. 7. – С. 46-
*diff*lib(python), Mercurial, Revlog

не вполне адекватны

Основные математические модели:

Широко используются (по Baudiš P. Current concepts in version control systems //arXiv preprint arXiv:1405.3496. – 2014.(2009-09-11)):

1. Longest Common Subsequence LCS (Myers E.: An $O(ND)$ Difference Algorithm and its Variations, *Algorithmica* 1 (1986) 251–266) — GNU diffutils, perl, ...
2. patiencediff (LCS for lines which occur exactly once on both sides, then recurse between). Cohen B., et al.: Bazaar Source Code, bzrlib/patiencediff.py, <http://www.bazaar-vcs.org/Download>
3. Bdiff longest common continuous substring — Ratcliff J. W., Metzener D. E. Pattern-matching-the gestalt approach //Dr Dobbs Journal. – 1988. – Т. 13. – №. 7. – С. 46-
difflib(python), Mercurial, Revlog

не вполне адекватны

Неадекватность модели LCS

~~определение~~ перемещения ~~я~~ сбережений

практика математика ~~практика~~

Удалено:

< о
< п

Добавлено:

> е
> ре

Изменено:

< Д
< м
> е

Изменено:

< л
< щ
> ение

Изменено:

< п
< сб
> ере

Изменено:

< меш
< ж
> ени

Изменено:

< я
< й
> й

Строка: с 1 по 1

Изменено:

< пр

▼ матем

а

Удалено:

< к

тика

Изменено:

< матем

▼ пр

а

Добавлено:

▼ к

тик

Строка: с 1 по 1

Удалено:

< ab cd ef gh ij kl mn op qr st uv e.g.
<

..

Добавлено:

> ab cd ef gh ij kl mn op qr st uv e.g.
>

Неадекватность patience diff

~~...~~ ab cd ef gh ij kl mn op qr st uv ~~e.g.~~ e.g. . .

Неадекватность Bdiff

~~...~~ ab cd ef gh ij kl mn op qr st uv ~~e.g.~~ e.g. . .

Smith, Temple F.; and Waterman, Michael S. (1981).
«Identification of Common Molecular Subsequences». Journal of
Molecular Biology 147: 195-197.
DOI:10.1016/0022-2836(81)90087-5. PMID 7265238.