パターンマッチング入門

2016年6月20日
明治大学・理工学部・情報科学科
笹尾　勤

パターンマッチングとは

- データを検索する場合に, 特定のパターンが出現するかどうか, またどこに出現するかを特定すること
- 応用
  - 検索
  - インターネット
  - アンチウイルス・ソフト
  - 迷惑メイル

3種のパターンマッチング

- 厳密マッチ(Exact match)
  - ビットパターンが完全に一致したものを検出
  - 端末アクセス制御装置で使用
- 正規表現(Regular expression)マッチ
  - (0 | 1)*000 (0 | 1)*
  - 迷惑メイル, ウイルス検出で使用
- 近似マッチング(Approximate match)
  - 最も似たパターンを探す

厳密マッチング

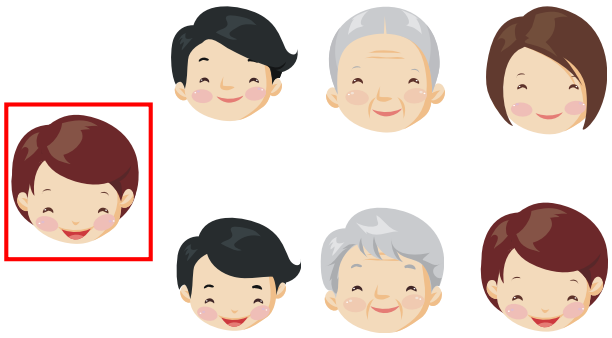
- 完全に一致しているパターンを見つける

101101

完全に一致

101101

同じ顔はどれか？



迷惑メイル

- viagra
- Viagra
- VIagra
- VIAgra
- VIAGra
- VIAGRa
- VIAGRA

正規表現マッチング

- VIAGRA
- VIAGRa
- VIAGra
- viagra

全部で64通り存在する.

(V|v)(I|i)(A|a)(G|g)(R|r)(A|a)

正規表現を使えば1行で表現可能

正規表現

- 文字列パターンの表記法.
- 通常の文字と, メタ文字と呼ばれる特別な意味を持った記号を組み合わせて利用.
- 文字列を直接指定せず, 特徴パターンを指定できる.

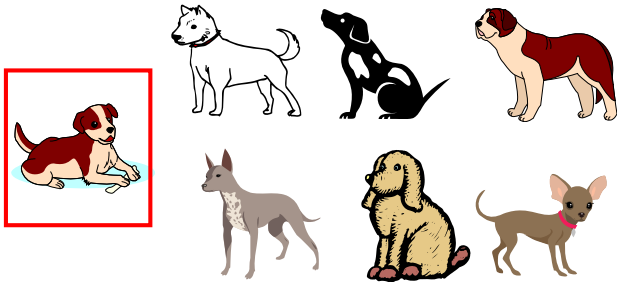
迷惑メイル

- viaagra
- viaaagraa
- viiagrra
- vviagra
- $v^+i^+a^+g^+r^+a^+$

近似マッチング

- 最も似ているパターンを見つける

最も似ている犬はどれか？



近似マッチング

- 最も似ているパターンを見つける
- ハミング距離：異なっているビット数

		ハミング距離
	110011	4
101101	110110	4
	101001	1

もとパターンとは緑のビットで異なっている

近似マッチング

- スペルチェック
  - univercity

    university
- DNAのパターン
  - 欠損を認める

完全一致の検出法

問題

- subway という単語の意味を知るために単語帳を参照する
- 英文字6個の単語帳には, 単語が804語ある.
- 前から順に率直に調べると, 最悪804回参照する必要がある.

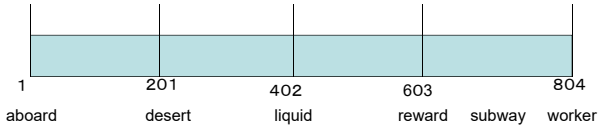
文字数6の単語帳(全部で804語)

- aboard abroad abrupt absent absorb accent accept
- access accuse across acting action active actual adjust
- admire adverb advice advise affair affect afford afraid
- agency agenda almost alumni always amount animal
- annual answer anyhow anyone anyway appeal
- appear around arrest arrive artist ashore asleep
- aspect assign assist assume assure atomic attach
- attack attend august author autumn avenue awaken
- ballet banana barber barely barrel basket battle beauty
- became become before behave behind belief belong
- beside better beyond biased bikini bishop bitter blonde
- bloody blouse boiler border boring borrow bother bottle
- bottom bought bounce branch breast breath breeze bridge
- bright broken bronze .....

高速な方法

- 英文字を2進数に置き換える.
- a は00001, zは11010
- 全ての単語を2進数に置き換える.
- 単語帳を予め, 数字の小さい順に並べ換える.

最悪でも10回の参照で検出できる



もっと高速な方法

専用回路を使う

通常のメモリ

1
2
3
4

情報科学科事務室
理工学部事務室
明治大学本部
小田急生田駅

ビット数が比較的少ない

ビット数が多い

連想メモリ

情報科学科事務室
理工学部事務室
明治大学本部
小田急生田駅

1
2
3
4

ビット数が多い

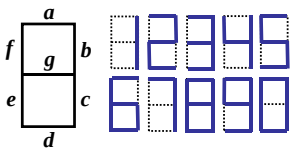
ビット数が少ない

連想メモリは
率直に作ると，非常に高価になる

安くつくる方法は？

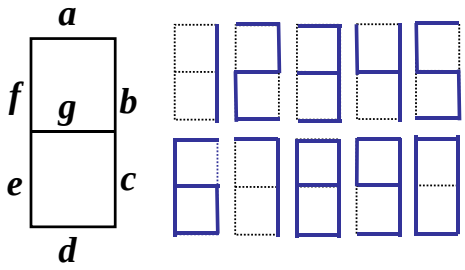
例：7セグメント　BCD変換回路

7セグメント							BCD
a	***b***	***c***	***d***	***e***	***f***	***g***	
0	1	1	0	0	0	0	1
1	1	0	1	1	0	1	2
1	1	1	1	0	0	1	3
0	1	1	0	0	1	1	4
1	0	1	1	0	1	1	5
1	0	1	1	1	1	1	6
1	1	1	0	0	**0**	0	7
1	1	1	1	1	1	1	8
1	1	1	1	0	1	1	9
1	1	1	1	1	1	0	A

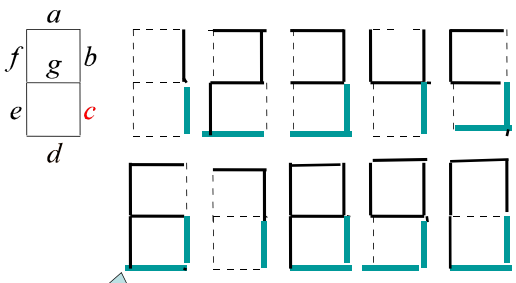


率直に作ると，
7入力のメモリが必要

数字の区別に必要なグメントは？



cとd以外のセグメント



5入力のメモリで十分

14文字からなる英単語を15個含む電子単語帳を作りたい.

なるべく小型に作る方法は？

14文字からなる英単語の表

- accommodations
- characteristic
- constitutional
- discrimination
- identification
- recommendation
- representative
- transportation
- administration
- congratulation
- disappointment
- generalization
- interpretation
- representation
- responsibility

メモリで率直につくると

- 英文字　26文字　5ビット
- 14文字:5x14＝70
- 70入力のメモリ:
- 2の10乗　1Kilo
- 2の20乗　1Mega
- 2の30乗　1Giga
- 2の40乗　1Tera
- 2の50乗　1Peta
- 2の60乗　1Exa
- 2の70乗　1Zeta
- 10の3乗
- 10の6乗
- 10の9乗
- 10の12乗
- 10の15乗
- 10の18乗
- 10の21乗 ← とんでもない数

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	f
a	c	c	o	m	m	o	d	a	t	i	o	n	s	1
a	d	m	i	n	i	s	t	r	a	t	i	o	n	2
c	h	a	r	a	c	t	e	r	i	s	t	i	c	3
c	o	n	g	r	a	t	u	l	a	t	i	o	n	4
c	o	n	s	t	i	t	u	t	i	o	n	a	l	5
d	i	s	a	p	p	o	i	n	t	m	e	n	t	6
d	i	s	c	r	i	m	i	n	a	t	i	o	n	7
g	e	n	e	r	a	l	i	z	a	t	i	o	n	8
i	d	e	n	t	i	f	i	c	a	t	i	o	n	9

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	f
i	n	t	e	r	p	r	e	t	a	t	i	o	n	10
r	e	c	o	m	m	e	n	d	a	t	i	o	n	11
r	e	p	r	e	s	e	n	t	a	t	i	o	n	12
r	e	p	r	e	s	e	n	t	a	t	i	v	e	13
r	e	s	p	o	n	s	i	b	i	l	i	t	y	14
t	r	a	n	s	p	o	r	t	a	t	i	o	n	15

Q: 英単語の区別に必要な文字数は?

A: 3文字.

x_3, x_6, x_{13}

5x3=15 ビットあればよい.

$$2^{15} = 32768 = 32Kilo$$

DNAのパターンマッチング

- DNAは4つのシンボルで表現されている
  - Adenine (アデニン)
  - Cytosine (シトシン)
  - Guanine(グアニン)
  - Thymine(サイミン)
- 高速パタンーマッチング回路を作る

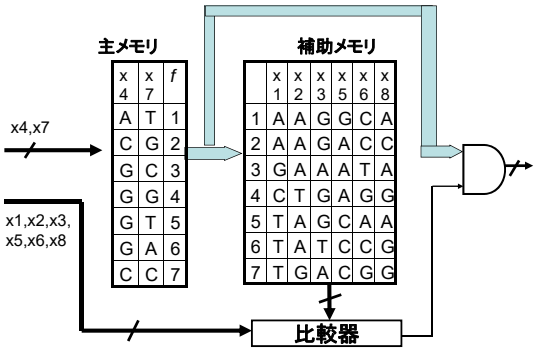
DNAパターンの検出回路

登録ベクトル								インデックス
x1	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*f*
A	A	G	A	G	C	T	A	1
A	A	G	C	A	C	G	C	2
G	A	A	G	A	T	C	A	3
C	T	G	G	A	G	G	G	4
T	A	G	G	G	A	T	A	5
T	A	T	G	C	C	A	G	6
T	G	A	C	C	G	C	G	7

DNAパターンの検出回路

登録ベクトル								インデックス
x1	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*f*
A	A	G	**A**	G	C	**T**	A	1
A	A	G	**C**	A	C	**G**	C	2
G	A	A	**G**	A	T	**C**	A	3
C	T	G	**G**	A	G	**G**	G	4
T	A	G	**G**	G	A	**T**	A	5
T	A	T	**G**	C	C	**A**	G	6
T	G	A	**C**	C	G	**C**	G	7

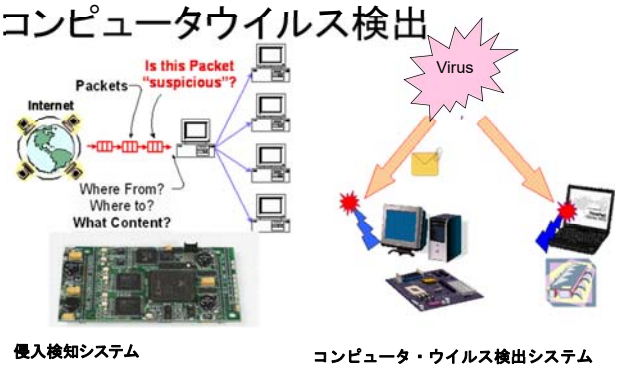
DNAマッチング回路



メモリ量の削減

- 率直な方法
  - 入力数 8x2＝16 ビット
  - 出力数 3ビット
  - メモリ量 $3x2^{16}=65536x3=196,608$
- 新しい方法
  - 主メモリ
    - 入力数 2x2＝4ビット
    - 出力数 8 ビット
  - 補助メモリ
    - 入力数 3ビット
    - 出力数 12 ビット
  - 総メモリ量$3x2^4+12x2^3=144$

コンピュータウイルス検出



侵入検知システム　　コンピュータ・ウイルス検出システム

コンピュータウイルス検出エンジン

- パターン数：130万
- 検出速度：毎秒3.2Giga パターン
- 通常のコンピュータよりも，千倍以上速い
- ウイルスは，毎月増えている
- 書き換えが可能

ウイルスパターンの例

- DOS.Trivial.27.J=b44ecd21ba????b43dcd2193b213
- HHH.1=50b9fb0f8b1e010181c3150180370043e2fa
- INF.Autorun-28=7368656c6c657865637574653d72656379636c65645c*2e657865

FPGAとSRAMで構成した コンピュータウイルス検出エンジン

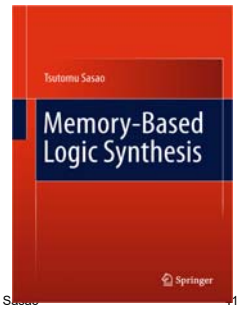
参考文献

- 丸山正明
- **産学官連携 大学がつくり出す近未来**
- 日経BP出版センター
- 2009/12/14

インターネットのウイルスチェック装置
九州工業大学の笹尾勤教授チーム
「高速パターンマッチング回路の合成と

参考文献

T. Sasao, “Memory-Based Logic Synthesis,” Springer, March 2011, pp.1-190.



日常生活でのパターンマッチング

- お札の判定
  - 色、印刷
  - 磁気インク
- コインの判定
  - 磁性体の強弱

生体認証

- 身体的特徴
  - 指紋
  - 静脈
  - 虹彩
  - 声紋
- 行動的特徴
  - 筆跡
  - まばたき