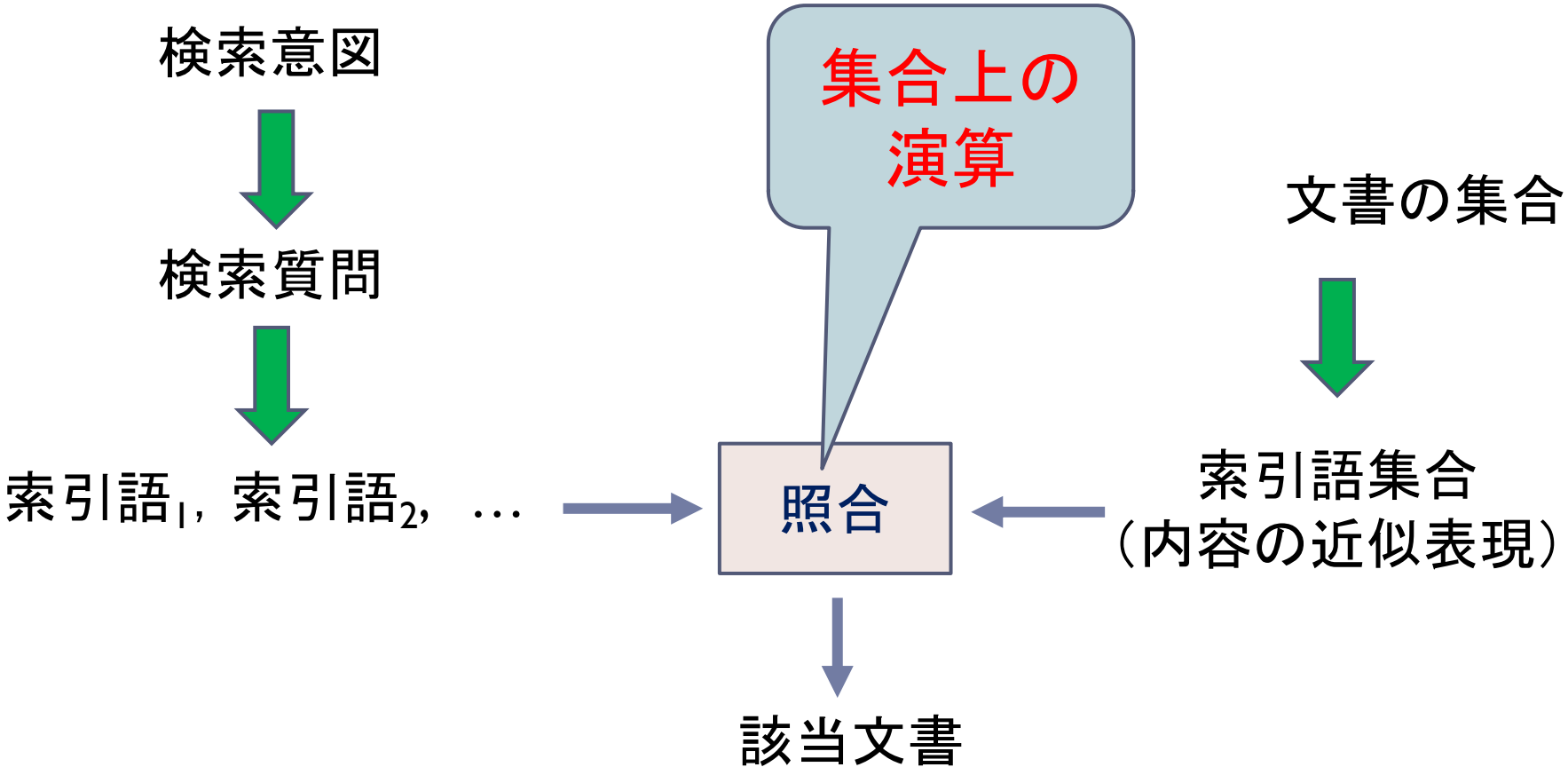
自然言語処理システム

（1）情報検索

情報検索の概念



転置インデックス法

- 検索質問(検索語の集合):T_1, T_2, …

  $T_1 \wedge T_2$:語T_1と語T_2の両方が含まれる文書を求める

  $T_1 \vee T_2$:語T_1または語T_2のどちらかが含まれる文書を求める

  ～T_1:語T_1を含まない文書を求める

- 例
  - 検索意図:「文法の学習に関する書籍や論文を探す」
  - 検索質問:文法∧学習
  - 検索意図:「英語以外の言語に対する文脈依存文法や文脈自由文法の学習に関するもの」
  - 検索質問:(～英語)∧(文脈依存文法∨文脈自由文法)∧学習

転置インデックス

文書と索引語

	索引語1	索引語2	索引語3	索引語4
文書1	1	1	1	0
文書2	0	1	1	1
文書3	1	0	1	1
文書4	0	0	1	1

転置インデックス

	文書1	文書2	文書3	文書4
索引語1	1	0	1	0
索引語2	1	1	0	0
索引語3	1	1	1	1
索引語4	0	1	1	1

検索の例

索引語1∧索引語2

索引語1	1010	= {文書1、文書3}
索引語2	1100	= {文書1、文書2}
索引語1∧索引語2	1000	= {文書1}

(索引語1∨索引語2)∧～検索語4

索引語1	1010	= {文書1、文書3}
索引語2	1100	= {書1、文書2}
索引語1∨索引語2	1110	= {文書1、文書2、文書3}
索引語4	0111	= {文書2、文書3、文書4}
～検索語4	1000	= {文書1}
(索引語1∨索引語2)∧～検索語4	1000	= {文書1}

転置インデックス法の拡張

文書と重みづけられた索引語

	索引語1	索引語2	索引語3	索引語4
文書1	0.2	0.5	0.6	0
文書2	0	0.3	0.1	0.8
文書3	0.5	0	0.5	0.2
文書4	0	0	0.3	0.3

重み付けられた索引語を利用した検索

	索引語2		索引語3			順位
文書1	0.5	+	0.6	=	1.1	1
文書2	0.3	+	0.1	=	0.4	3
文書3	0	+	0.5	=	0.5	2
文書4	0	+	0.3	=	0.3	4

ベクトル空間法

1. 文書と検索質問の両方を統一的に表現する。
2. この間で、距離（類似度）を定義し、似ている文書を探し出す。

文書をベクトルの線形結合で表したもの：

$D_r = \sum_{i=1}^{t} a_i^r V_i$

V_i：検索語T_iに対応するベクトル

a_i^r：文書D_rにおける索引語T_iに対する値

- 例）D_rにT_iが存在すれば1、otherwise 0
- 例）索引語T_iの重要度

検索質問をベクトルの線形結合で表したもの：

$$Q_s = \sum_{i=1}^{t} a_i^s V_i$$

類似度

内積:$|D_r||Q_s|\cos\alpha$

$V_i \cdot V_j$=1 …i=j
$V_i \cdot V_j$=0 …i≠j

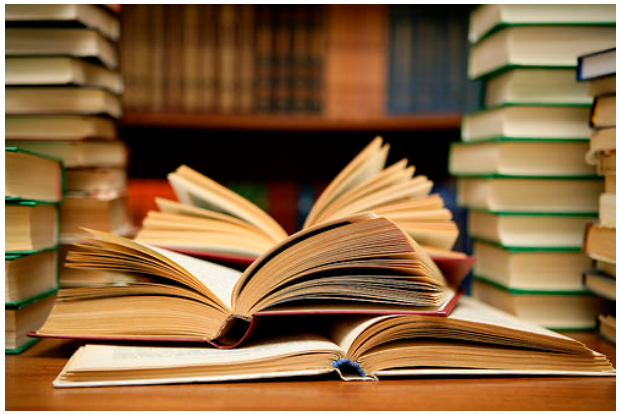
$$\mathrm{sim}(Dr,Qs) = Dr \cdot Qs$$

$$= \sum_{i,j=1}^{t} a_i^r \, a_j^s V_i \cdot V_j$$

$$= \sum_{i=1}^{t} a_i^r \, a_i^s$$

実際の検索

1. あらかじめ各文書に対する文書ベクトルを計算しておく。
2. 検索質問を、検索質問ベクトルに変換する。
3. 検索質問と全ての文書ベクトルの類似度を計算する。
4. 類似度の大きい順にソートする。
5. 上位M位までの文書を出力する。



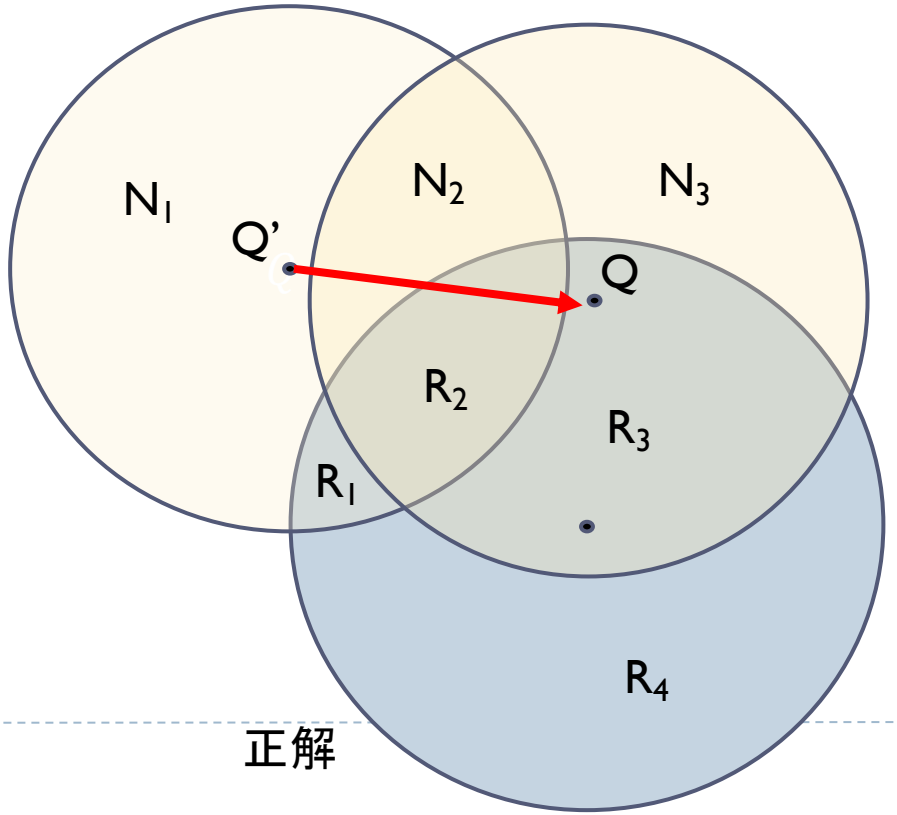
例

- 文書（D_1～D_3）と検索質問のベクトル表現
  - $D_1 = 3V_1 + 2V_2 + 4V_3 + 0V_4$
  - $D_2 = 1V_1 + 3V_2 + 0V_3 + 2V_4$
  - $D_3 = 2V_1 + 4V_2 + 1V_3 + 5V_4$
  - $Q = 1V_1 + 0V_2 + 2V_3 + 0V_4$
- 類似度計算
  - $sim(D_1,Q) = 3・1+2・0+4・2+0・0=11$
  - $sim(D_2,Q) = 1・1+3・0+0・2+2・0=1$
  - $sim(D_3,Q) = 2・1+4・0+1・2+5・0=4$
- 結果
  - 順位１：D_1（類似度１１）、順位２：D_3（類似度４）、順位３：D_2（類似度１）

関連フィードバック法

- 質問ベクトルの変更（質問Q'を、より適切な質問Qに変更する）

$$Q = Q' + \frac{1}{|R|}\sum_{D_i \in R} D_i - \frac{1}{|N|}\sum_{D_j \in N} D_j$$



N1:含まれなくなった不正解
N2:含まれる不正解
N3:新しく含まれる不正解
R1:含まれなくなった正解
R2:含まれる正解
R3:新しく含まれる正解
R4:含まれない正解

判定結果に対する評価
再現率(Recall Ratio)と適合率(Precision Ratio)

- 再現率

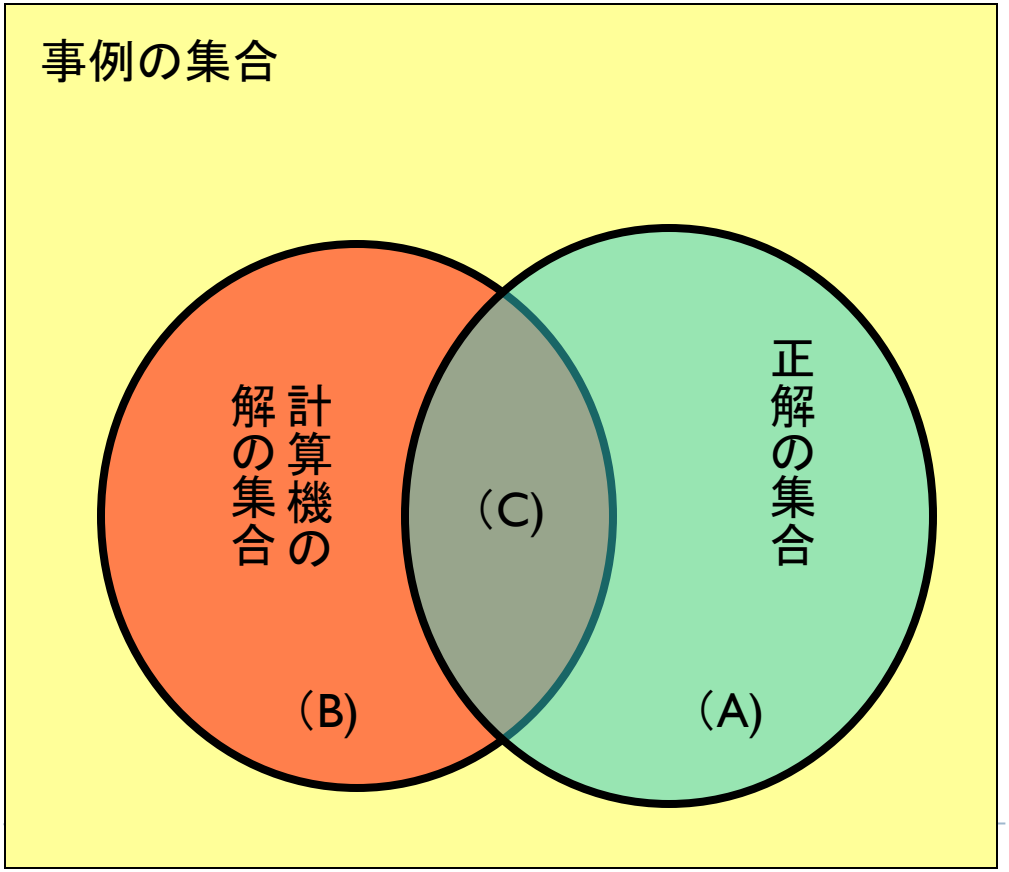
$$R = \frac{|C|}{|A|+|C|}$$

- 適合率

$$P = \frac{|C|}{|B|+|C|}$$

- F値

$$F = \frac{RP}{R+P}$$



重要語句の抽出

- 基本的には出現頻度(tf. term frequency)が多い語が重要。
- 語wの（一文章中の）頻度： tf(w)
  - しかし、頻出の語でも多くの文章に出現するものと、特定の文章だけに出現する語では重要さに差がある。

idf(inverse document frequency)

- 出現の偏りを表すための指標
- 特定の文章に出現する頻出語（一般的でない語）ほど重要であると考えられる。
- 語wのidf値：　idf(w) = - log(n/N)
  - nは文章集合（文章数N）のなかで語wが含まれる文章数。
  - n/Nは、文章中に語wが現れる生起確率
- これを重みとして出現頻度に乗じたものが

  tf(w)*idf(w)

  値である。
- 文章集合として、資料の文章全体を使う。

全文検索

1. Text

$$S = s_1 s_2 s_3 \cdots s_n$$

s_i : 文字

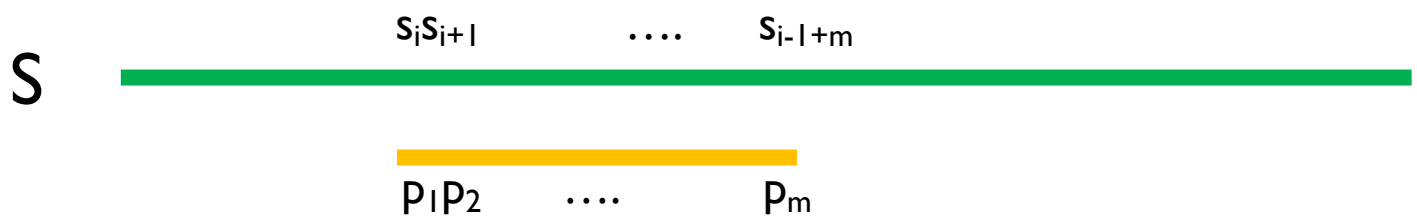
2. 検索する文字列

$$P = p_1 p_2 p_3 \cdots p_m \quad (m \leqq n)$$

p_i : 文字

3. 目的： iを見つける

$$s_{i-1+k} = p_k \quad (k=1,\ldots,m)$$



アルゴリズム

```
begin
   for i := 1 until n - m + 1 do
      begin
         j := 1;
         while (p[j] = s[i + j - 1]) do
            begin
                  if (j = m) then terminate with    result i
               elsif (j < m) then j := j + 1;
            end;
      end;
end;
```

効率化：

- Knuth-Morris-Prattのアルゴリズム
- Boyer-Moorのアルゴリズム

テキストの分類（階層的クラスタリング）

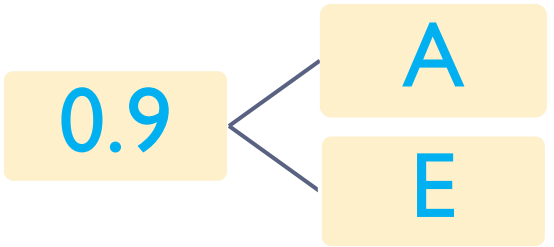
1. 各テキストを一つずつクラスタとする
2. クラスタが一つになるまで、次を繰り返す
   1. それぞれのクラスタ間の類似度を計算する
   2. 最も類似度の高いクラスタの組を一つのクラスタに併合する

1.

	A	B	C	D	E
A	.	.3	.6	.8	.9
B	.3	.	.5	.7	.8
C	.6	.5	.	.4	.1
D	.8	.7	.4	.	.3
E	.9	.8	.1	.3	.

併合
A - E 0.9

階層木



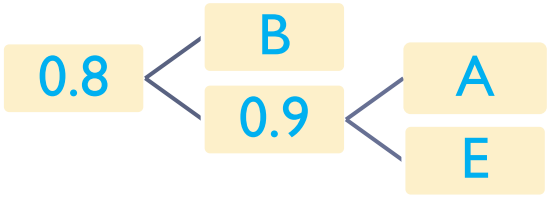
2.

	AE	B	C	D
AE	.	.8	.6	.8
B	.8	.	.5	.7
C	.6	.5	.	.4
D	.8	.7	.4	.

併合

AE - B 0.8

階層木



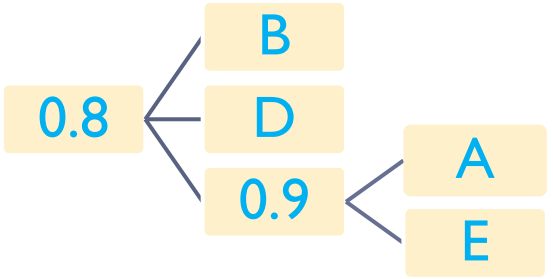
3.

	ABE	C	D
ABE	.	.6	.8
C	.6	.	.4
D	.8	.4	.

併合

ABE - D 0.8

階層木

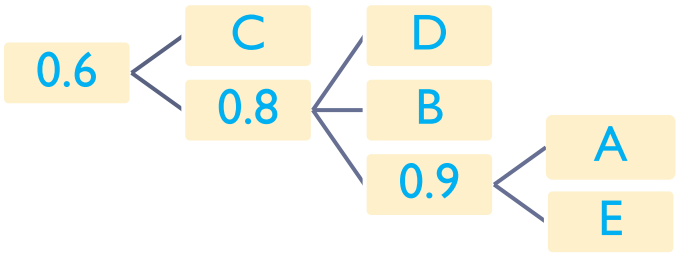


4.

	ABDE	C
ABDE	.	.6
C	.6	.

併合
ABDE - C 0.6

階層木



クラスタ間の類似度の考え方

1. クラスタX、クラスタYの要素中の最も大きいもの

$$\text{sim}(X,Y) = \max_{x \in X, y \in Y} (sim(x,y))$$

2.クラスタX、クラスタYの要素中の最も小さいもの

$$\text{sim}(X,Y) = \min_{x \in X, y \in Y} (sim(x,y))$$

3.クラスタX、クラスタYの要素の平均値

$$\text{sim}(X,Y) = \text{average}_{x \in X, y \in Y} (sim(x,y))$$

テキストの要約

- 抽出した情報をテキスト（文章）で表現する。

- 理論的（理想的）には、

  文章 ⇒ 理解 ⇒ 再構成 ⇒ 文章生成

- 理解：
  - 重要な部分の同定
    - 文章構造
      - 序論、本論、結論
    - 重要度の高い文を残す
      - 重要度の計算

パラメータ
（1） キーワードの出現回数
（2） 特定の表現パターンの存在
（3） 時制（過去、現在）
（4） 文のタイプ（主張、推測、事実、etc）
（5） 前文との接続関係（理由、例示、逆説、並列、対比、接続、etc）
（6） 文章中の位置
（7） 段落中の位置

まとめ

- 情報検索
  - 転置インデックス法
  - 同（重み付き）
  - ベクトル空間法
  - 関連フィードバック法
- 判定結果の評価
  - 再現率、適合率
- 重要語句の抽出
- 全文検索
- テキストの分類
- テキストの要約

課題

文書（D_1～D_3）に対して、

- $D_1 = 3V_1 + 2V_2 + 4V_3 + 0V_4$
- $D_2 = 1V_1 + 3V_2 + 0V_3 + 2V_4$
- $D_3 = 2V_1 + 4V_2 + 1V_3 + 5V_4$

次の検索質問のそれぞれについて、各文書（D_1～D_3）との類似度を求めよ

1. $Q_1 = 0V_1 + 2V_2 + 2V_3 + 0V_4$
2. $Q_2 = 1V_1 + 0V_2 + 0V_3 + 2V_4$
3. $Q_3 = 0V_1 + 2V_2 + 0V_3 + 1V_4$