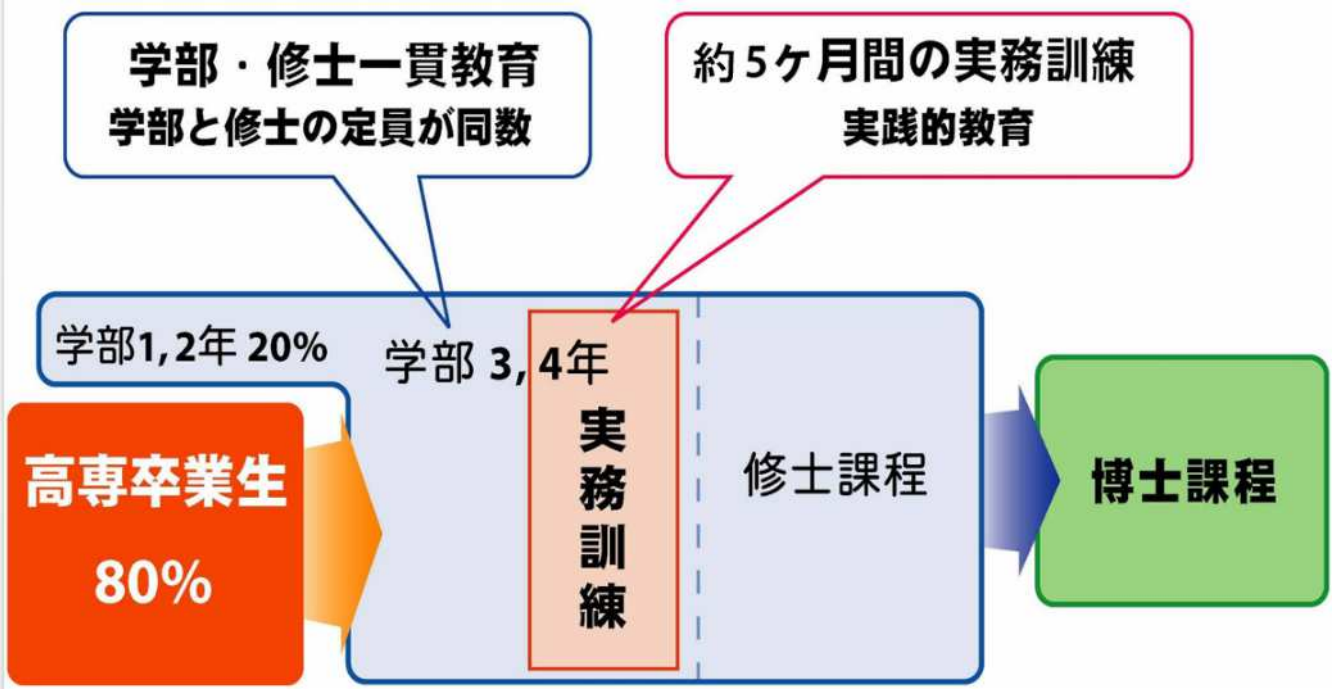
テキストマイニング技術とその 応用

産学官協同における

長岡技術科学大学
経営情報系 （来年度より「情報・経営システム工学専攻」）
湯川 高志

長岡技術科学大学のご紹介

- 高専のための大学です



長岡技術科学大学のご紹介（つづき）

- 開学（1976年）当初から，企業等との共同研究を積極的に推進
- 実践的・指導的技術者の育成
  - 「技術科学=技術を科学的態度でブレークスルーする」ことができる技術者を育成
- 学部4年生の後半の実務訓練
  - 約5ヶ月間，実社会において，身につけた知識や技能を応用する訓練（単なる「体験」とは異なる）
  - 15%程度の学生は海外において実務訓練を実施
- 積極的な国際交流
  - 約12%が留学生
  - 約90の国際交流協定締結機関

テキストマイニング技術

テキストマイニングとは

- テキスト　　マイニング

文字で（人間が読むために）書かれた情報

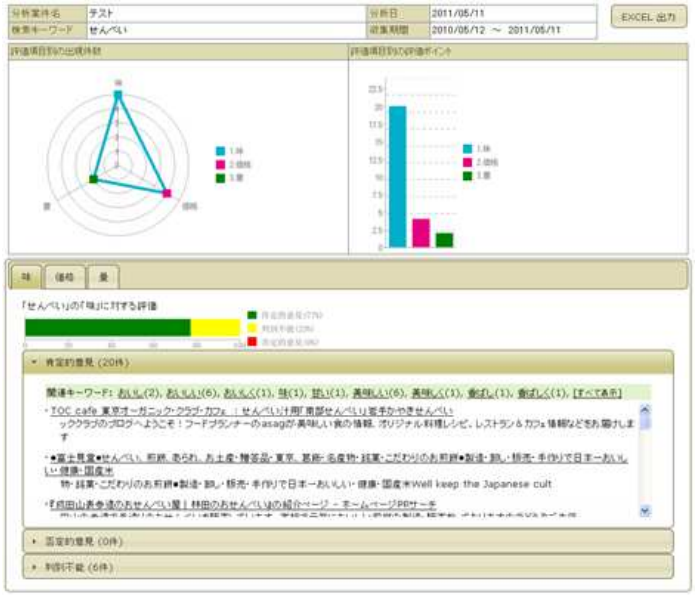
発掘

人間が読むために文字で書かれた文章（自然言語）に対し，単語の出現頻度，類語関係，相関関係などを分析して，有用な情報を取り出すこと

- 情報検索とは違うのか？
  - 情報検索: 欲しい情報を（ひとつ）見つける
  - テキストマイニング: 関連する情報の集合から知見を得る

テキストマイニングが使われる分野（の例）

- CRM (Customer Relation Management)
  - コールセンタやコンタクトセンタに寄せられる顧客からの問合せや苦情を分析し，問題解決法を迅速に提示したり，製品の問題の早期発見，製品改良などにつなげる
- 評判分析
  - ブログや電子掲示板（，最近ではTwitterも）から，製品やサービスに対する口コミ情報を抽出して，評判を分析し，製品の改良や新規開発につなげる



テキストマイニングが使われる分野（の例）（つづき）

- 特許情報処理
  - 特許文書（特許明細書）の記述を分析し
  - 技術開発トレンドや相互関係を「見える化」する（特許マップ）
  - ある特許を無効にする特許をさがす

解決すべき問題

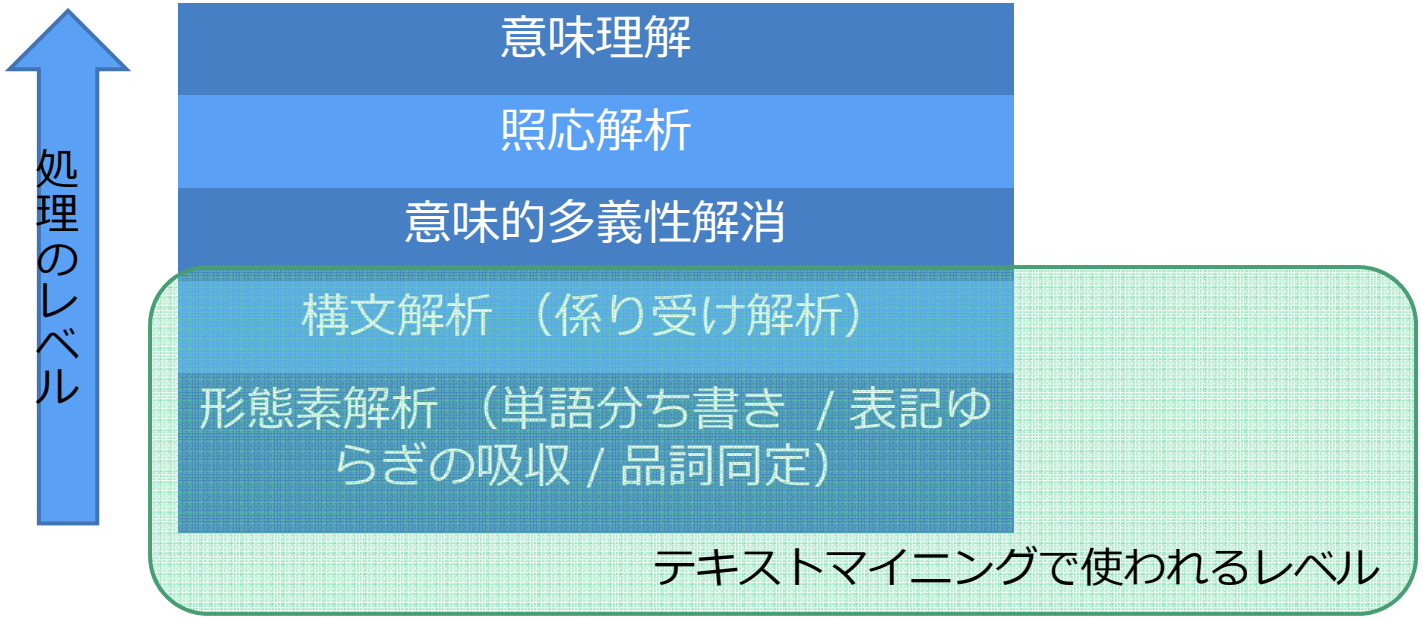
解決手段		電気的			光学的	
		電流密度増大	過電流保護	消費電力の低減	放出の安定性	放出の強度
結晶構造	多結晶			1998-145000 1998-256782		1998-012917 1998-093141
	アモルファス			1998-165702 1998-302514	1998-256158 1998-294515	1998-172086
電極材料	単一材料	1998-242517	1998-215034	1998-241518 1998-279213 1998-310236	1998-163527	1998-145002 1998-200162
	合金	1998-135516 1998-247761	1998-223930 1998-294489	1998-135514 1998-179827 1998-290543 1998-344764	1998-150021 1998-190049 1998-244774	1998-012933 1998-256597

テキストマイニングが使われる分野（の例）（つづき）

- 社会情報処理
  - ネットワークにある（テキストに限らない）情報から，人間関係を推定する
    - 「あの人検索 スパイシー」（オーマ株式会社）
- 推薦システム
  - その人のSNSへの書込みや「いいね」に基づいて興味を抽出し，それに合ったニュースを提示する
    - 「Gunosy」（株式会社Gunosy）

テキストマイニングに使われる技術（1）

- 対象とするテキストは人間が読むための言葉（自然言語）で書かれているため，自然言語処理が必要



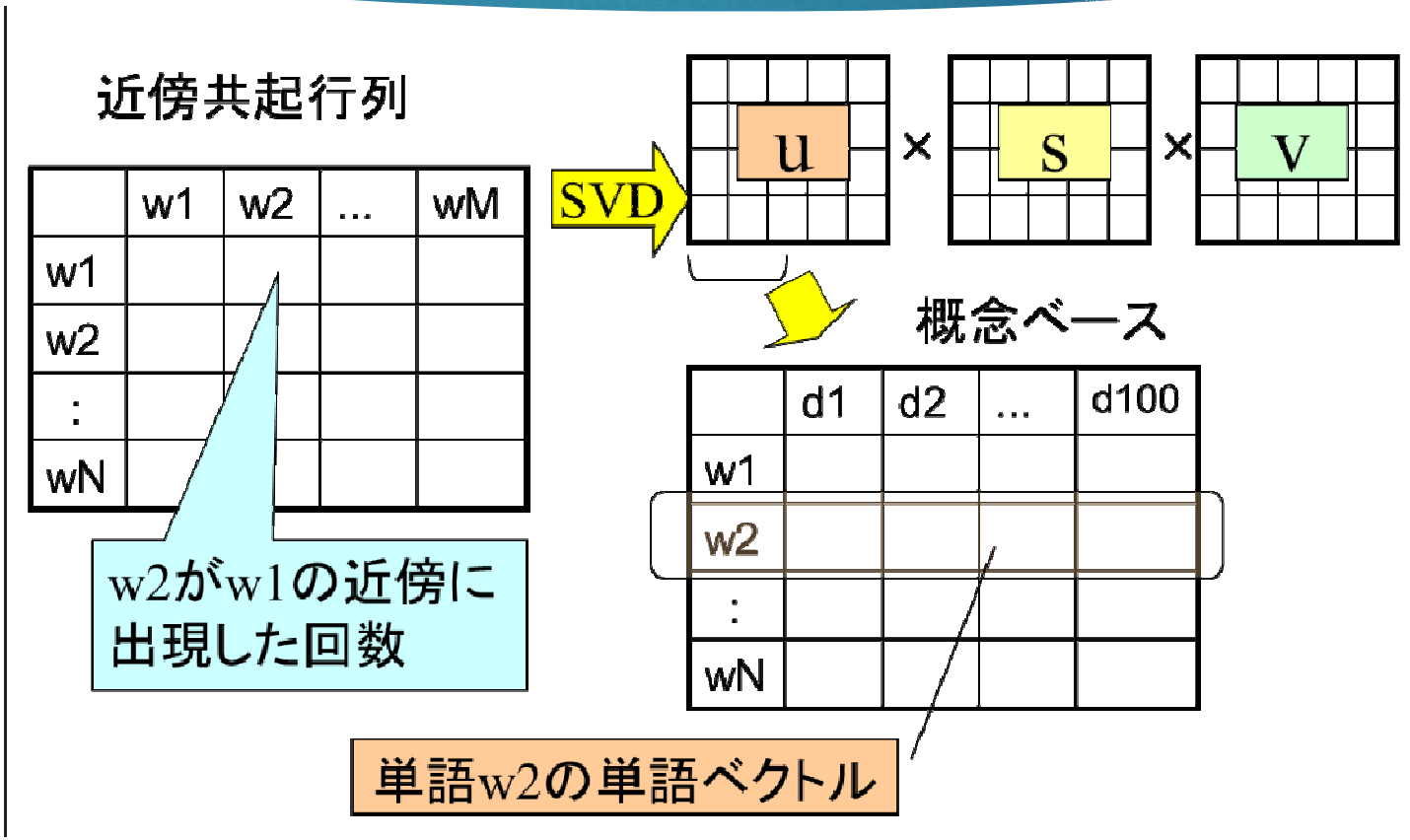
テキストマイニングに使われる技術 (2)

- 語の意味的類似度の判別 – 「概念ベース」
  - 語Aと語Bについて
    この2語だけが違い他は同じ，という文が多くある場合
    語Aと語Bとは似ているだろう
  - 例
    - 「近年の**コンピュータ**は，計算性能が大幅に向上するとともに，低価格化している」
    - 「近年の**電子計算機**は，計算性能が大幅に向上するとともに，低価格化している」

  - 文章の中で語がどのような語の近くに現れるかに着目

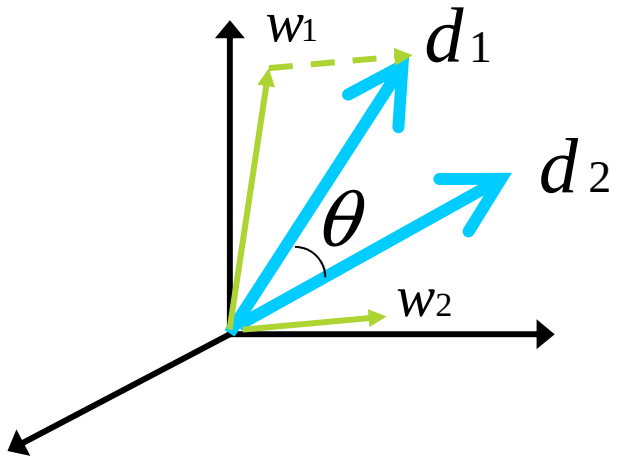
近傍共起

概念ベースの構築方法


近傍共起行列
w1
w2
...
wM
w1
w2
:
wN
w2がw1の近傍に
出現した回数
SVD
u
×
S
×
V
概念ベース
d1
d2
...
d100
w1
w2
:
wN
単語w2の単語ベクトル

概念ベースを用いた意味的類似判別

- 文や文書=語の集合 (bag of words)
  - 文書のベクトル
    =（文書に含まれる語のベクトルの総和）を正規化
- 語も文書も同一のベクトル空間に配置される
- 文書や語の間の類似度=ベクトルのなす角の余弦値
  - 語 vs 語
  - 語 vs 文書
  - （語, 語, ．．．） vs 文書
  - 文書 vs 文書



$$R(d_1, d_2) = \cos\theta$$

分析手法 （1）

- 文書や単語のベクトルが作れると．．．

- 意味の類似性に基づく検索
- 多次元→2次元のマッピングにより，単語や文書の関係を「見える化」
- 単語や文書の似たもの同士のまとまりを作る=「クラスタリング」
- ベクトルに基づいて，どちらに分類されるかを判別する=「クラシフィケーション」
  - 分類の境界を求めるのに，機械学習が使われる場合も多い