情報探索入門（第3回）

分類の一般概念と分類理論

京都大学情報学研究科
黒橋禎夫
kuro@i.kyoto-u.ac.jp

分類の一般概念と分類理論

- 「分類は知のはじまり」
- 物事を体系化→全体を把握

- 分類 (Classification)
- 類似性 (Similarity)

目 次

分類の演習

なす、新聞、ほうき、キカイダー、にわとり、

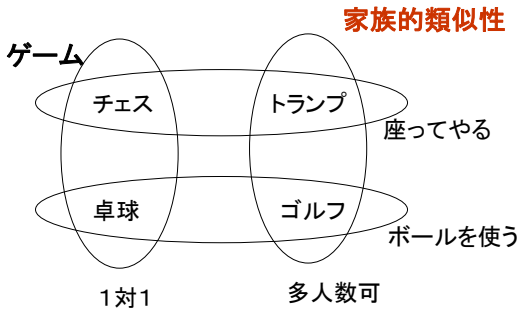
リンゴ、学生、いす、トマト、コンピュータ、

ピラニア、テレビ、掃除機、くじら

視点･観点

分類は、視点･観点によって異なる

ウィトゲンシュタイン



言葉、文化との関係

- 言葉⇔概念
  - 山 ： 高くもりあがった地形
  - 平野 ： たいらに広がった地形
  - 丘 ： ？
- 文化
  - ドイツではトマトは果物
  - 日本での魚の細かい名前

オーバーゾーニング

- 百貨店の売り場
  - 地下：食品、1階：化粧品、2階：洋服
    3階：スポーツ用品、…
- オーバーゾーニング
  - スキーの売り場：スキー用品、ツアー予約、
    チェーン、道路地図、健康飲料、…

- いす、くじら、なす、にわとり、ほうき
- キカイダー、コンピュータ、テレビ、トマト、
  ピラニア、リンゴ
- 学生、新聞、掃除機

動植物の分類

動植物の分類

- アリストテレスの動物分類
  - 血液の有無、生殖のタイプ、足の数
  - 人為分類
- 17世紀　航海技術の進歩、珍しい動植物
- リンネ（分類学の父）の動植物分類
  - 階層的カテゴリ
  - 名前を属名と種名で表す

階層的カテゴリ

界　動物界
門　脊髄動物門
綱　哺乳綱
目　食肉目
科　イヌ科
属　イヌ属
種　イヌ種

- アダンソンの植物分類
  - 多くの形質を考慮し、多くを共有するものをグループ化
  - 類型分類
- ラマルクの動物分類
  - 動物の進化の系統を再現する分類
  - 系統分類
  - ダーウィンの「種の起源」後、盛んに研究
    - 化石などでわかることは小数
    - 形態学的、発生学的、細胞学的形質による類型分類

図書の分類

図書館の歴史

- 古代
  - アレキサンドリア図書館、蔵書目録
- 中世
  - 修道院や教会の図書館
  - 数百から2000冊程度
- ルネッサンス以降
  - 大学、学問分野、主題による分類

図書館の歴史

- 18世紀
  - 教育、中産階級
  - 会員制図書館、貸本屋
- 19世紀～
  - 公共図書館
  - 十進分類法、コロン分類法

図書の分類

- 書架分類
  - 図書館の棚のどこに何をおくか
- 書誌分類
  - 書誌情報（タイトル、著者名、主題等）の分類
  - 主題の分類を設定
  - そこへ各図書を対応付ける

十進分類法（デューイ、国際、日本）

000 総記	700 芸術
100 哲学と心理学	710 生活、造園
200 宗教	720 建築学
300 社会科学	730 造形美術、彫刻
400 言語	740 絵画、装飾美術
500 自然科学と数学	750 画法、絵
600 技術（応用科学）	760 工芸美術、印刷、版画
700 芸術	770 写真術、写真
800 文学と修辞学	780 音楽
900 地理学と歴史	790 娯楽、演芸

コロン分類法

40ほどの主題を設定

z	総記	BZ	物理的科学
1	知識	C	物理学
2	図書館学	D	工学
3	図書学	E	化学
4	ジャーナリズム	F	技術
A	自然科学	G	生物学
AZ	数理科学	H	地学
B	数学	…	…

コロン分類法（ファセット）

- 医学
  - 器官 ： 眼、胃、血液、骨、…
  - 分科 ： 解剖学、生理学、疾病、衛生、…
- 絵画
  - 様式 ： 日本画、西洋画、宗教画、…
  - 素材 ： 人物、風景、静物、…
  - 材料 ： 紙、木、ガラス、…
  - 技法 ： 構図、色彩、水彩、油絵、…

ことばの分類

シソーラス
（語を体系的に整理したもの）

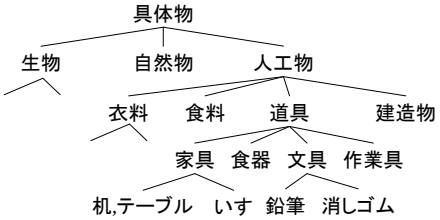
- 一般用語については、上位下位よりも同義語関係が中心
- 単語の選択の手助け

ex. 角川類語新辞典
分類語彙表（国立国語研）
ロジェのシソーラス
Longman Language Activator

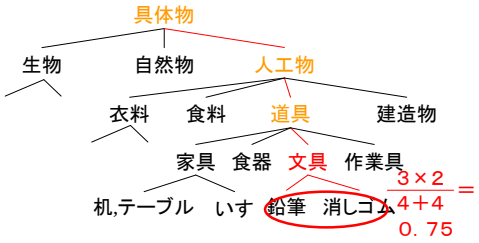
専門用語のシソーラス

- 分野の学問体系を明らかにする
  （専門用語集＋α）
- 文献検索での統制言語
  - 等価関係（優先語、非優先語）
  - 階層関係（上位語、下位語）
  - 連想関係

自然言語処理でのシソーラスの利用

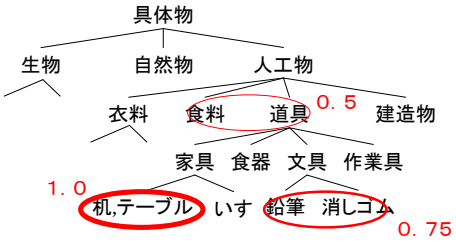


自然言語処理でのシソーラスの利用



$$類似度＝\frac{(共通の親の深さ)\times 2}{各語の深さの和}$$

自然言語処理でのシソーラスの利用



$$類似度＝\frac{(共通の親の深さ)\times 2}{各語の深さの和}$$

用例ベース翻訳

- 女性洋服売り場はどこですか。
- 婦人服売り場はどこですか。
  Where can I find ladies dresses?

分類の数学的理論

分類の数学的理論

- 人為分類　：　少数の形質を人為的に選択
- 類型分類　：　多くの形質の共有を調べる
  （アダンソンの植物分類）
  → クラスター分析などの数量分類学

数量分類学

- 特徴ベクトル（属性の束）で個体を表現
- 個体間の類似度＝特徴ベクトルの類似度
  – 一致係数、ユークリッド距離、角度
- クラスタ分析
  – 類似度の高いものをまとめる

特徴ベクトル

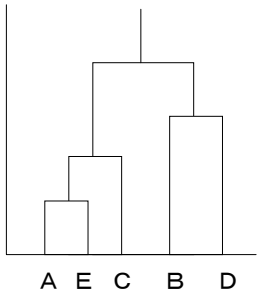
属性

個体	f1	f2	f3	f4	f5	f6
A	0	1	0	0	1	1
B	1	0	1	1	1	0
C	0	1	0	1	0	0
D	1	0	1	0	0	1
E	0	1	0	1	1	1

類似度（一致係数）

	A	B	C	D	E
A	1	1/6	3/6	2/6	5/6
B		1	2/6	3/6	2/6
C			1	1/6	4/6
D				1	1/6
E					1

クラスター分析



樹形図

情報検索

情報検索

テキストの特徴ベクトル表現→類似度計算

- 図書検索
- 新聞記事検索
- 電子メール検索
- WWWページ検索

インターネット

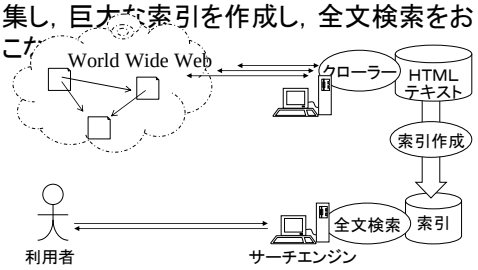
- 広義：複数のコンピュータネットワークの相互接続
- 狭義：国際的に広く相互接続されたもの（The Internet）
- 歴史：
  1969年 アメリカの国防総省によるARPANET
  1984年 日本の学術組織の研究用ネットワークJUNET
  1991年 欧州素粒子物理学研究所のティム・バーナーズ=リーがWorld Wide Webプロジェクトを発表
- 特定の集中した責任主体はなく，接続している組織が各ネットワークを管理

ウェブ（World Wide Web, WWW）

- インターネット上で提供されるハイパーテキストシステム
- 文書はHTML（ハイパーテキスト記述言語）で記述，別文書への参照（リンク）を埋め込むことでインターネット上の文書の相互参照を可能とする
- ウェブディレクトリ：1994年 Yahoo!
- 検索エンジン
  1994年 WebCrawler, Infoseek, Lycos
  1995年 AltaVista, Excite（日本：Yahho, ODiN, 千里眼）
  1998年 Google
- 文書数
  - 英語（”the”のヒット数）：116億
  - 日本語（”の”のヒット数）：19億

検索エンジン＝クローラー＋全文検索

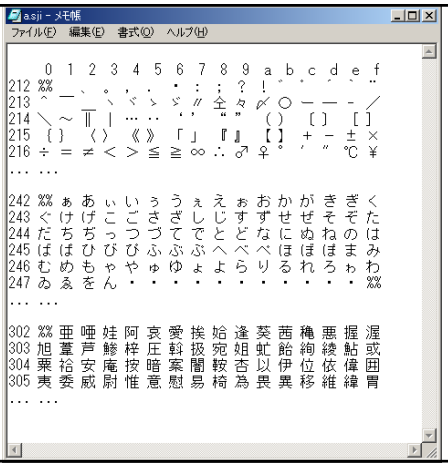
- ハイパーリンクをたどってHTML文書を収集し，巨大な索引を作成し，全文検索をおこな



文字コード

- 文字 a ⇔ 文字コード 01100001

- 文字コードセット：ある特定の文字集合
  - ASCIIコードセット（ISO8859－1）
  - JISローマ字（JIS X0201 ローマ字）
  - JIS漢字（JIS X0208）



辞書式順序

- 文字間の順序関係：文字コード順
  - A < B < C
  - B < a
- 二つの文字列の順序関係：
  はじめにあらわれた異なる文字間の順序
  - AB < ABC < ABD
  - かかく < かかし < かがく

転置インデックス（索引）

文書1	言語、コンピュータ、問題
文書2	コンピュータ、問題
文書3	言語、問題、情報
文書4	問題、情報

言語	文書1、文書3
コンピュータ	文書1、文書2
問題	文書1、文書2、文書3、文書4
情報	文書2、文書3、文書4

語の重要度（TF.IDF）

語の頻度（Term Frequency）

TF	文書1	文書2	文書3	文書4	IDF
言語	2	0	1	0	2
コンピュータ	1	1	0	0	2
問題	2	2	3	1	1
情報	0	1	2	1	1.3

全文書数 / 語の出現する文書数
（Inverse Document Frequency）

語の重要度（TF.IDF）

言語 問題　検索

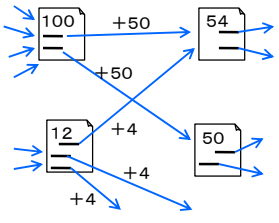
TF.IDF	文書1	文書2	文書3	文書4
言語	4	0	2	0
コンピュータ	2	2	0	0
問題	2	2	3	1
情報	0	1.3	2.6	1.3
	6	(2)	5	(1)

PageRank

- 「多くの良質なWebページから参照されている
  Webページは良質である」

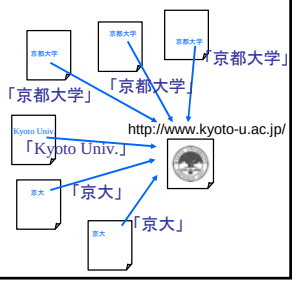
$$R(u) = \sum_{v \to u} \frac{R(v)}{|Bv|}$$

$$\mathbf{R} = c\mathbf{AR}$$



アンカーテキストの利用

- アンカーテキスト：リンクが張られた文字列
  例：京都大学
- アンカーテキストはリンク先テキストの一部とみなす
  - 特定のトピックに関連し、被参照数の大きいWebページが検索されやすい
  - リンク先に含まれない語句でも検索できる（例："京大"）



ま と め

- 分類 ⇔ 類似性
- 動植物分類の歴史
  - 人為分類、類型分類、系統分類
- 図書の分類法
  - 十進分類法、コロン分類法
- ことばの分類
  - シソーラス
- 数量分類，情報検索

次回（5/1）：演習

- 場所：メディアセンター203、204
  注意：メディアセンターのアカウントを確認しておくこと！！
- E-Learningシステムを利用予定
  https://cms.ecs.kyoto-u.ac.jp/
  利用マニュアル
  http://www.iimc.kyoto-u.ac.jp/services/ecs/WebCT/PDF/Student-Tool_Ver1.pdf
- 演習課題：
  - 書籍のNDC分類、KULINEの利用
  - ウェブページの分類