情報探索入門（第2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

（2011年10月17日）

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)
- 分類法・学 (taxonomy)
- 類似性 (similarity)

目 次

分類の演習

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、

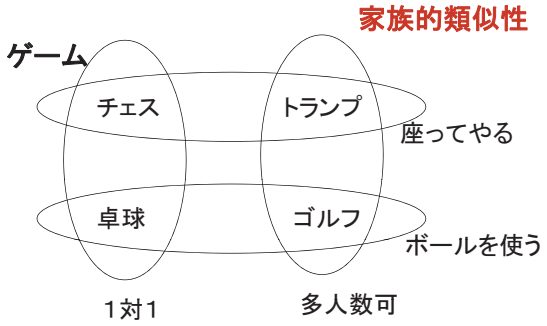
ピラニア、テレビ、掃除機、くじら

分類の諸問題

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

リンネ博物館（ストックホルム）

- 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）

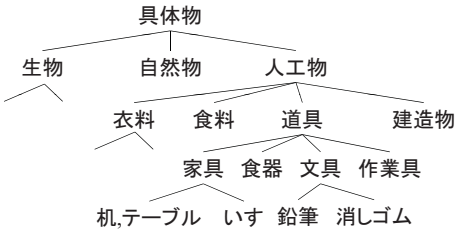
- 一般用語については、上位下位よりも同義語関係が中心
- 単語の選択の手助け

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

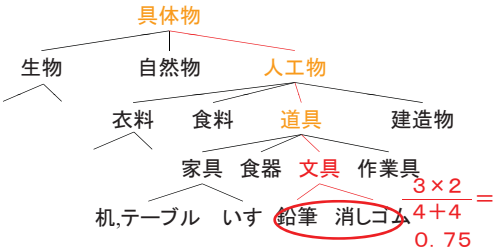
専門用語のシソーラス

- 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

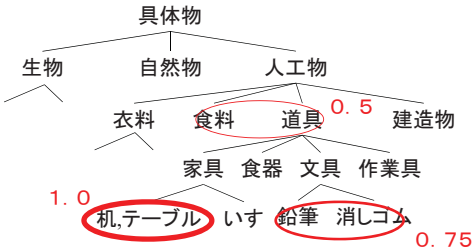


自然言語処理でのシソーラスの利用



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

自然言語処理でのシソーラスの利用



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

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

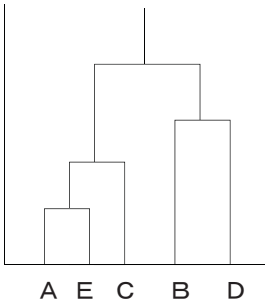
属性

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

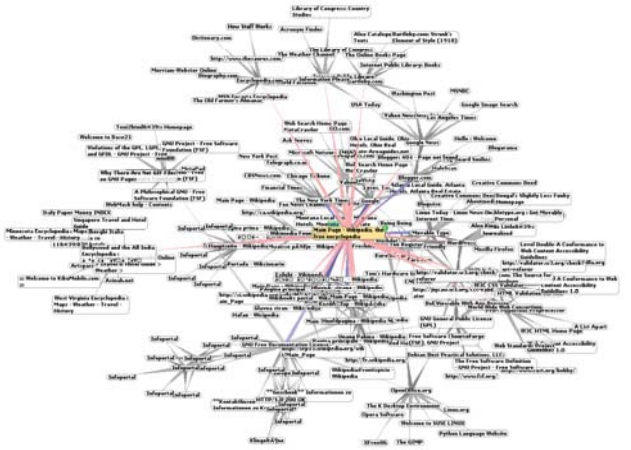
インターネット

- 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

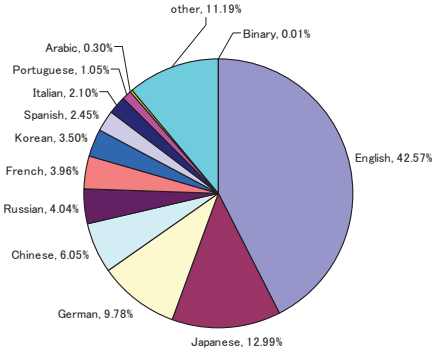
  1984年 日本の学術組織の研究用ネットワークJUNET

  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



出典：http://ja.wikipedia.org/wiki/Www

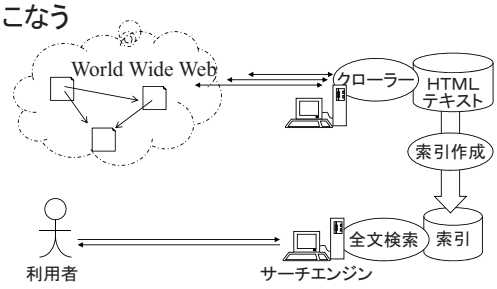
ウェブページ（107億）の言語分布



出典:童芳，平手勇宇，山名早人: 全世界のWebサイトの言語分布と日本語を含むWebサイトのリンク・地理的位置の解析, DEWS2008.

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，全文検索をおこなう



転置インデックス（索引）

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

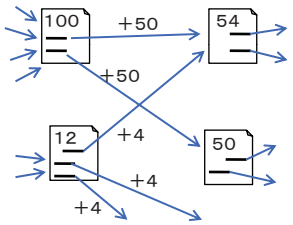
TF.IDF	文書1	文書2	文書3	文書4
言語	4	0	2	0
コンピュータ	2	2	0	0
問題	2	2	3	1
情報	0	1.3	2.6	1.3
	6	(2)	5	(1)

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」

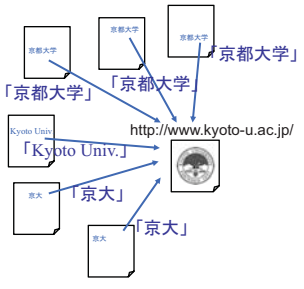
$$R(u) = \sum_{v \to u} \frac{R(v)}{|Bv|}$$

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



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"）



情報の信頼性

- 玉石混交のウェブ
- 多様な視点からの分類・分析結果の提供

フォークソノミー (folksonomy)

- folks（民衆）＋taxonomy（分類法）
- ユーザによるウェブ上の情報へのタグ付け（分類）
- 共同作業による分類．タグの検索やタグを付けた人，その人がつけた他のタグを調べることができる．

例） はてなブックマーク（ソーシャルブックマーク），Flickr（写真共有サイト），ニコニコ動画（動画共有サイト）

ソーシャルブックマーク

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
- フォークソノミー，ソーシャルブックマーク

10/24,31：演習

- 場所：学術情報メディアセンター203、204
- 演習課題：
  - 書籍のNDC分類
  - KULINEの利用
  - ソーシャルブックマークの利用
- 準備
  - メディアセンターのアカウントの確認

情報探索入門（第3回）

分類演習1

10月24日（月）2限目

情報学研究科　黒橋禎夫教授

分類担当チーム　演習補助者

1

分類とは？（講義の復習）

- 分類　視点・観点によって異なる
  例）書店
  ジャンル：スポーツ、園芸、語学・・
  サイズ：文庫・新書コーナー
  時間：新刊コーナー
- 図書館の本は分類ごとに並べられている。
  日本十進分類法（NDC）がよく使われている。

2

十進分類法（デューイ、国際、日本）

- デューイ

000　総記
100　哲学
200　宗教
300　社会科学
400　語学
500　自然科学
600　応用科学（技術）
700　芸術、レクリエーション
800　文学
900　歴史

- 日本

000　総記
100　哲学
200　歴史
300　社会科学
400　自然科学
500　技術
600　産業
700　芸術
800　言語
900　文学

3

日本十進分類法（NDC）

000　総記
100　哲学・宗教
200　歴史
300　社会科学
400　自然科学 →
500　技術
600　産業
700　芸術
800　言語
900　文学

400　自然科学
410　数学
420　物理学
430　化学
440　天文学、宇宙科学
450　地球科学、地学
460　生物科学、一般生物学
470　植物学
480　動物学
490　医学、薬学

4

分類の利点・欠点

〇分野で探せる。同じテーマの資料が集中。
　関連したテーマも探せる。
〇タイトルにキーワードがなくても検索可能。
〇体系的・階層性がある。

×複数のテーマをもつ書籍も配架場所は1つに絞らなければならない。
×新しいテーマ・分野への対応が難しい。

5

分類を使う―文献収集

- レポート作成には参考資料が必要
- 先行研究を調べる文献収集が重要
- キーワード検索だけで済ませていませんか？
- 関係のある分類の書架にある本を眺めてみる「ブラウジング」をしてみよう！
  ～偶然の発見も文献探しでは大切～

6

ブラウジング

- キーワードが分からなくても、実際の書棚をうろうろする、実物を手に取る。
- 関連分野の本が一堂に会している。
- キーワード検索だけでは見つけられない新たな発見があるかもしれない。
- 思っているところにあるとは限らない。

7

演習の目的

- 演習A：分類検索を使ってみる。
  ブラウジングの効果を知る。
  キーワード検索では見つけられないこともあるということを知る。
- 演習B：書籍を分類してみる。
  分類すること、1つに絞ることの難しさを知る。
- 演習C：書籍以外を分類してみる。
  分類の利点・欠点を知る。

8

演習A（分類検索を使ってみる）

- ブラウジングの効果を知る。
- キーワード検索では見つけられないことも
  あるということを知る。

A-1

あなたは講義の予習のため図書館で本を探そうとしています。
どの棚に必要そうな本があるか、日本十進分類法(NDC)の3桁の分類番号で考えられるかぎり全て答えてください。

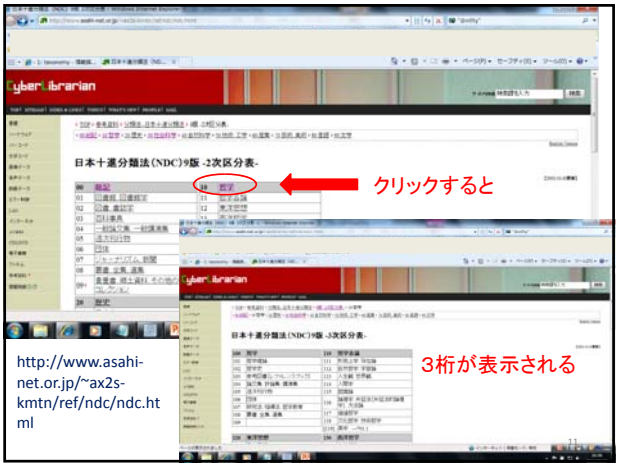
A-2

A-1で答えた分類番号でKULINEを分類検索し、検索結果のなかから《タイトルに「科学技術」あるいは「安全」という言葉を含まないが参考になりそうな書籍》のタイトルを1つ答えてください。

9

Open Course Ware

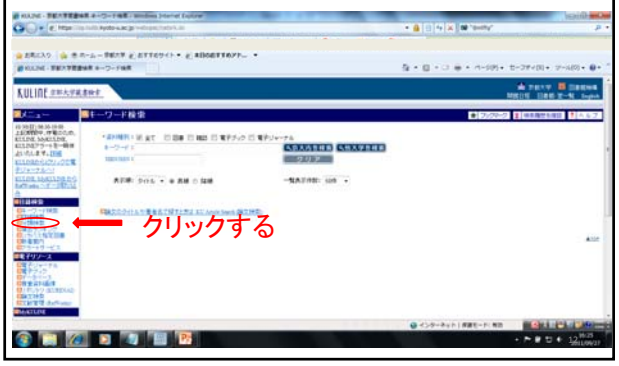
10

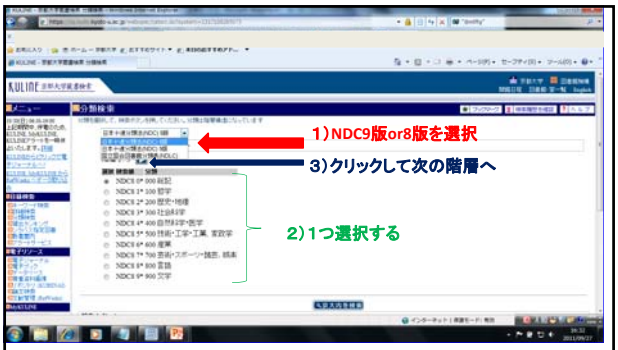


http://www.asahi-net.or.jp/~ax2s-kmtn/ref/ndc/ndc.html

11

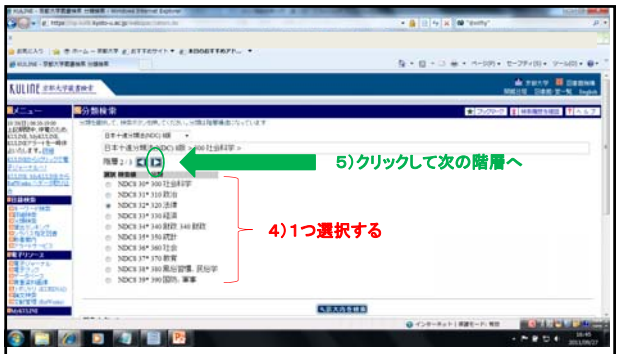
分類検索するには





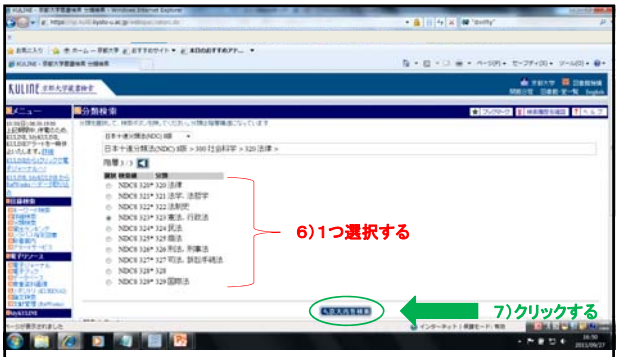
階層1/3

13



階層2/3

14



階層3/3

15

分類検索結果

演習B（書籍を分類する）

- 分類すること、1つに絞ることの難しさを知る。

B-1

以下の書籍から5冊選び、NDCに基づいて分類してください。分類番号(3桁)を考えられるかぎり全て答えること。

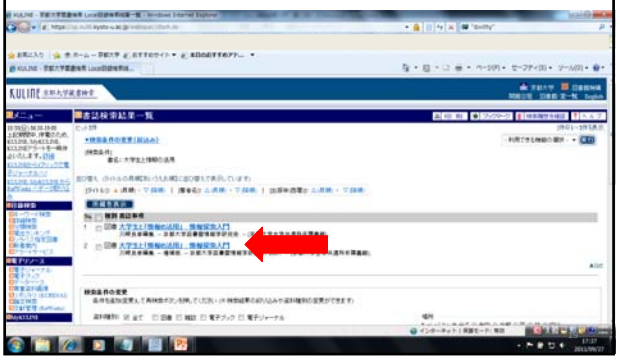
GoogleBooksで目次や本文の一部を読めるので参考にしてください。

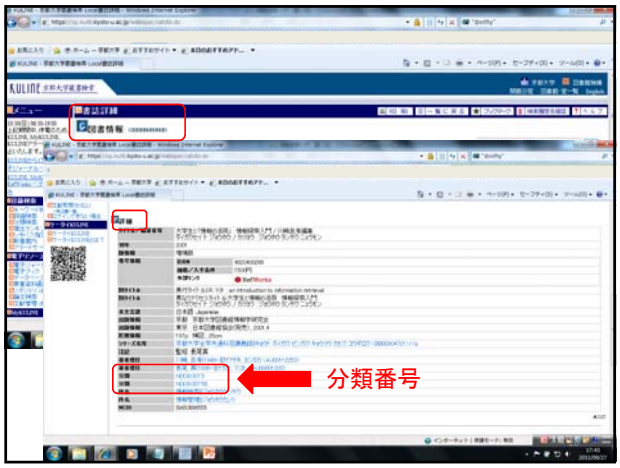
B-2

B-1で答えた分類番号を最も適したもの1つに絞り、それを選んだ理由を述べてください。またKULINEの分類と比較して相違点などを考察しなさい。

17

分類番号の表示は？





分類検索結果

演習C（書籍以外を分類してみる）

- 分類の利点・欠点を知る

以下の新聞や京大HP上のお知らせの見出しから4つ選び、NDCに基づいて分類してください。
それぞれ最も適した分類番号(3桁)を1つずつ答えること。
※見出しだけで分類を考えてください。

22

提出方法

「答案の雛形」をコピーして答案を作成し、メールで提出する。

- 提出先: ensyu@kulib.kyoto-u.ac.jp
- 件名:「情報探索入門分類演習1 氏名 学生番号」
- 締切【厳守】: **2011/10/31 (月) 12:00 [日本時間]**

- 答案はメール本文に直接書くこと (ファイルを添付しない)。
- 作成中の答案は消えてしまわないようにこまめに保存すること。
- 10/24の演習時間内に答案を提出しない方へ。PCのデスクトップに保存したファイルはログアウトすると消えるのでUSBメモリなどに保存すること。

23

質問・相談があれば補助者までどうぞ！

24

情報探索入門　第4回

分類演習2
10/31（月）2時限目

情報学研究科
黒橋禎夫教授
分類担当チーム

分類とは？

- 分類　視点･観点によって異なる
- 図書館の本は分類ごとに並べられている
- ウェブは？

フォークソノミー(folksonomy)

- folks(民衆)+taxonomy(分類法)
- ユーザーによるウェブ上の情報へのタグ付け（分類）
- 共同作業による分類。タグの検索やタグを付けた人、その人が付けた他のタグを調べることが出来る。

フォークソノミーの例

- Flickr（写真）
- ニコニコ動画、YouTube（動画）

フォークソノミーの例

- Citeulike(文献）
- Mendeley（文献）
- Connotea（文献）
- Delicious（web）

ソーシャルブックマークの特徴

- 手元のブックマークをインターネット上に保存することが出来る。
- 携帯やパソコンなどさまざまな環境で使える。
- タグ付けとコメントにより分類の手間を軽減
- ソーシャル機能
  - みんなからブックマークされている人気のサイトが分かる。
  - 同じ興味を持った人がブックマークしているサイトをチェックできる。

演習の目的

- フォークソノミーを知る･利用する
  - ウェブページを収集してブックマークし、タグやコメントを付与する。
  - 他の人の「分類」の仕方を参考にする。
  - 自分の「分類」について考察する。
  - タクソノミーとフォークソノミーの違いについて考察する。

はてなアカウントは取得していますか？

- 取得していない方は情報探索入門のページのはてなID取得方法を見て登録してください。
- 取得している方ははてなブックマークにログインしてください。
  - 学生番号やECS-IDをはてなIDとして登録してませんか？
- お気に入り機能を使用します。簡易ログインをされている方はログインしなおしてください。

はてなブックマークの画面

ブックマークレット

- はてなブックマークレット
  - http://b.hatena.ne.jp/register
  - ブックマークする Myブックマーク の両方を追加してください。
  - 右クリック→お気に入りに追加
  - セキュリティ警告：追加しているお気に入りは、安全でない可能性があります。続行しますか？→「はい」をクリック
  - 作成先：お気に入りバーに追加

ブックマークの方法

- ページを見つける
- お気に入りバーの ブックマークする をクリック
- コメントの部分に記入する。タグは[ ]で囲む。
  - [京都大学]でタグになる。
- ブックマークを確認をチェックして「追加する」をクリック

問題A

A-1.京都大学のHPのお知らせページ
[2011年度　ニュースインデックス (研究成果)]
http://www.kyoto-u.ac.jp/ja?type=monthly&c2=4
の中から、自分の興味のあるお知らせを自由に3つ選び、はてなブックマークに登録してください。
- 必ず複数のタグとコメントを付けてください。

A-2.他のユーザのつけたタグやコメントを確認し、自分のタグと同じ点、違う点について考察してください。
- 選択しているのが自分だけの場合、タグ付けで工夫した点を答えてください。

2011年度ニュースインデックス（研究成果）

他人のタグ･コメントを見る方法

他人のタグ･コメントを見る方法

問題B

- レポートを書くとして下記の3つのテーマから1つを選んでください。
  - a.災害とボランティア活動
  - b.再生可能エネルギー
  - c.ソーシャルメディアとアフリカ

B-1.選んだテーマに有用そうなページを検索エンジンで探して10個ブックマークし、あわせて必ずタグとコメントをつけてください。

B-2.自分がつけたタグと同じものがつけられているページや、同じページをブックマークした人がブックマークしている別のページから、有用そうなページを10個探し、ブックマークしてください。それぞれには必ずタグとコメントをつけてください。

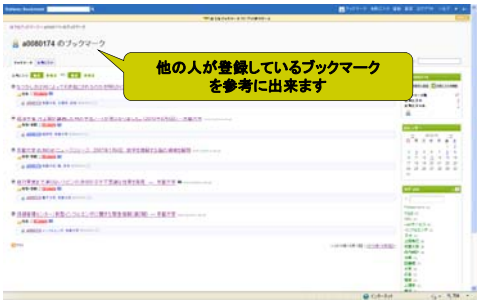
B-3.Webページの検索やタグ付けする際に工夫した点を自由に考察してください。

タグのリンク

タグのリンク

他の人のブックマークを見る方法

他の人のブックマークを見る方法



問題C

- 10/24の演習1のような分類（NDC）と今日の演習2のような分類（共同作業によるタグ付け）の違いについて、それぞれのメリット・デメリットを考察してください。

提出方法

- 「答案の雛形」をコピーして答案を作成し、メールで提出する。
- 提出先：ensyu@kulib.kyoto-u.ac.jp
- 件名：「情報探索入門分類演習2 氏名 学生番号」のようにすること
- 締切：**【厳守】2010/11/7（月） 12:00 [日本時間]**
- 答案には必ずはてなアカウント（ID）を記入すること。
- 答案はメール本文に直接書くこと（ファイル添付しない）。
- 作成中の答案は消えてしまわないようにこまめに保存すること。
- 10/31の演習時間中に答案を提出しない方へ。PCのデスクトップに保存したファイルはログアウトすると消えるのでUSBメモリなどに保存すること。

アンケート

- 分類の講義・演習（10/17,24,31）に関するアンケートにも答えてください。
- 演習問題ページからアンケートへリンクしています。
- 締切：2010/11/7（月） 12:00 [日本時間]