知識流通支援のツールとしての巨大検索ディレクトリ

鳥澤 健太郎

北陸先端大

自動生成された巨大ディレクトリ「鳥式」

- 科研特定領域「情報爆発IT基盤」

現代日本人のミッション

- 新しい価値あるものの発見、創出
- インターネットへの注目
  - 新しい価値あるものが「転がっている」
    - キャッチーなキーワード：ネット検索、ロングテール、テキストマイニング、ブログ、....
      - 全ては新規な価値あるもののソース！

知識を得る、伝える

- 興味/フォーカス
- 想定外
- 価値

知識流通

- 興味/フォーカス
  - 膨大な知識から簡単に興味あるものを発見/提示
- 想定外
  - これがなければ知識流通は成立せず
- 価値
  - 無価値な知識は流通させない

知識流通支援システム

- 興味/フォーカス、想定外
  - 知識の網羅的な索引付け （サーチ、自動生成）
- 価値
  - 多様な価値観を自動生成された索引付けにどう反映させるか？

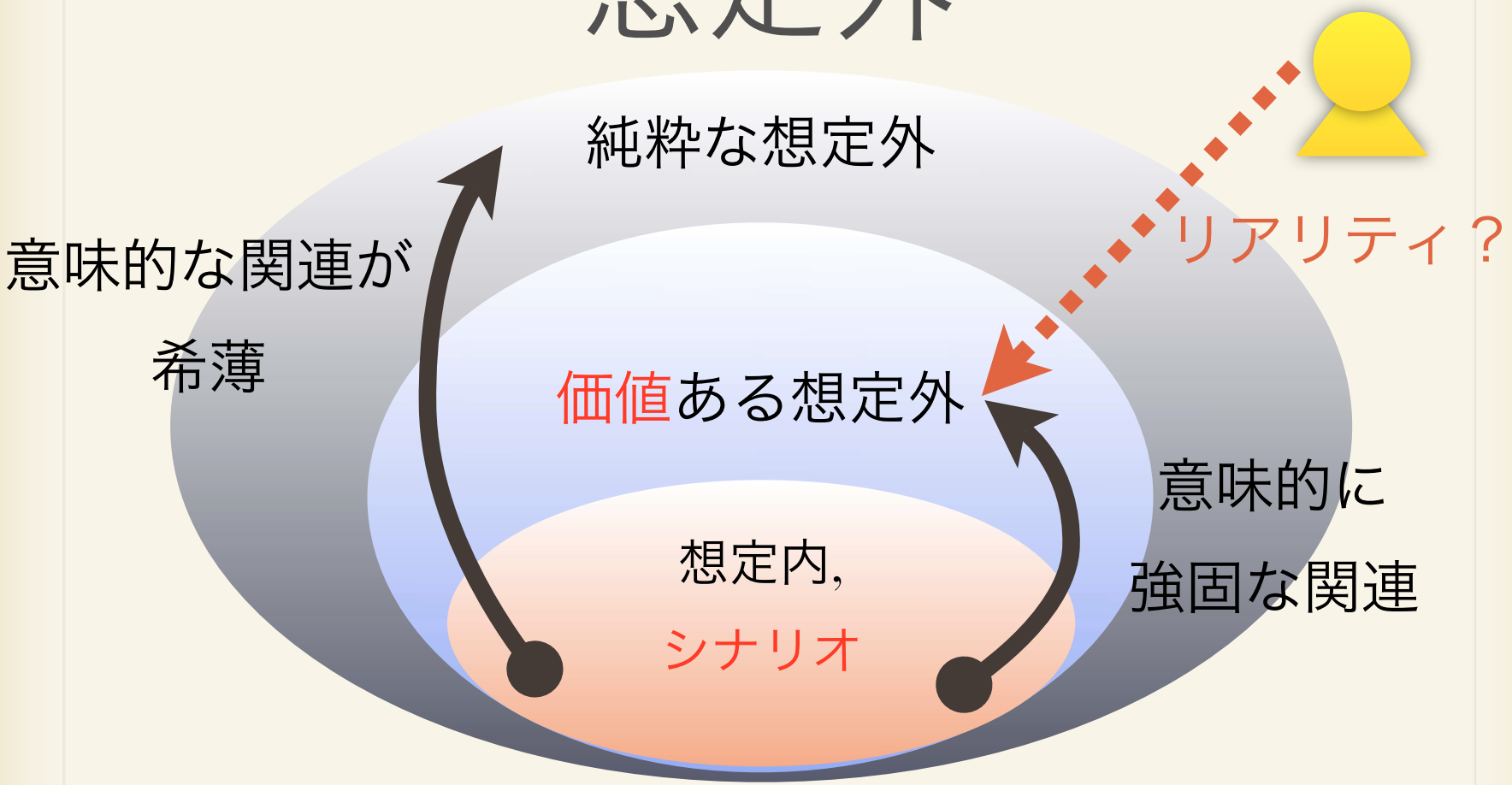
価値とはなにか？

- とんでもない難問
- ただし、価値とは「シナリオ」に埋め込まれて初めて定義される
  - 「砂漠の水一杯」？
  - 「価値とはなにか？」が難問なのは、シナリオを妥当な方法で限定できないが故
  - SNS⇒友達が言っていることには価値がある

目標&アプローチ

- 目標：Web上での「価値ある知識」の流通支援
  - 想定外でありながら価値あるものも提示
  - ロングテールもアクセス可能
- アプローチ
  - 大量のWeb文書からの知識の自動獲得
    - 価値を特徴づけるシナリオ（検索対象の利用）

想定外



想定外とは本来向こう側から勝手に到来するもの

⇒意識的なアクセスの必要な現在のサーチ技術ではなかなか見つけるのが難しい

想定外

- アオブダイ
- 外道としてアオブダイが釣れた
- 刺身が食べたい→”アオブダイ 刺身”で検索

想定外

- 検索結果
  - 大量のブログエントリ
    - うまそうな写真
    - 激賞の嵐
  - ある大学の海洋博物館：「食用」
- ところが....

想定外

- ”アオブダイ　毒”で検索すると、
  - 関西では中毒で、死者、入院が多数でている
  - 厚生労働省で販売禁止の通達
    - 業者の間でも周知がされていない旨警告した魚市場の広報紙

というわけで、

- アオブダイを検索している人にとって、トラブルとしての「毒」には価値がある
- こうした価値あるものを提示できるシステムがあればよい？

想定外

- 東京ディズニーランド

我が家のちびどもがウルトラマンからミッキーマウスに宗旨替えしたとすると、

想定外

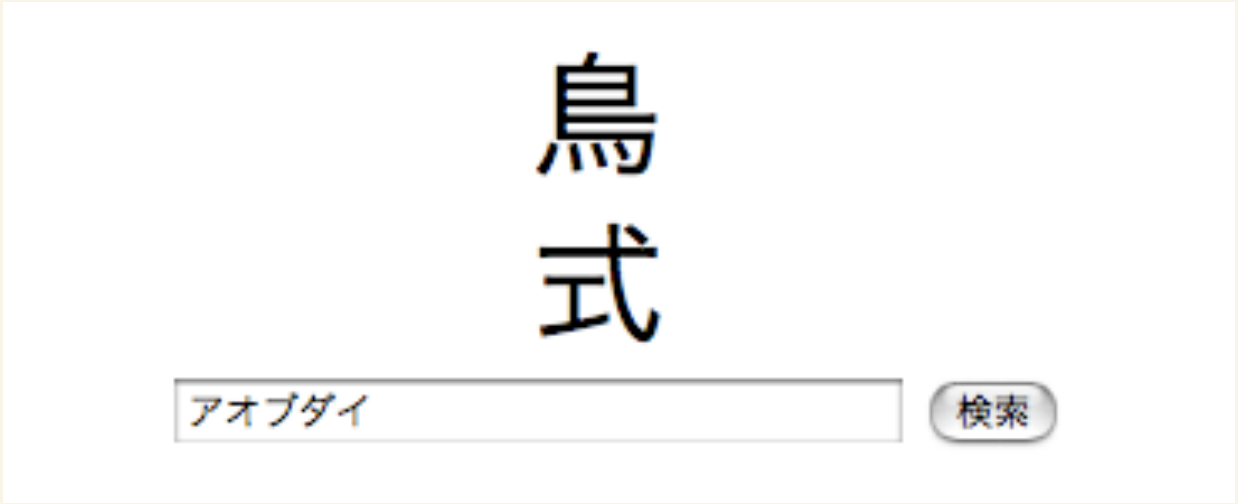
- 東京ディズニーランド: 父親にとって最悪の場所
  - 地雷が沢山
    - 身長制限: 身長が90cm以上ないとスプラッシュマウンテンに乗れない
    - 渋滞、混雑
  - 豆にチェックするほど暇な父親はあまりいない

鳥式ディレクトリ

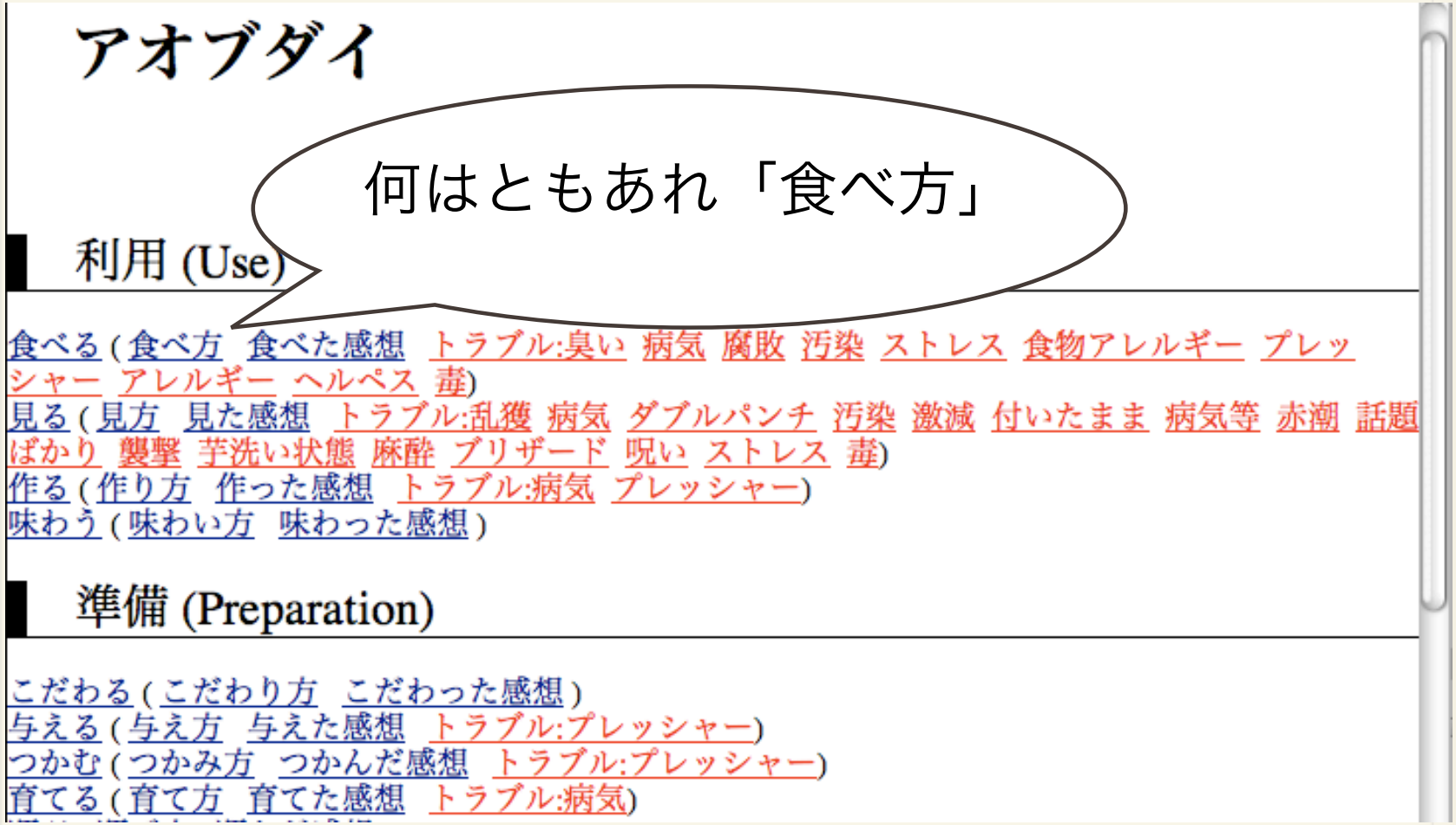
Webコーパスから獲得された

知識ベースに基づいて自動生成

アオブダイ



アオブダイ


アオブダイ
何はともあれ「食べ方」
利用 (Use)
食べる (食べ方 食べた感想 トラブル:臭い 病気 腐敗 汚染 ストレス 食物アレルギー プレッシャー アレルギー ヘルペス 毒)
見る (見方 見た感想 トラブル:乱獲 病気 ダブルパンチ 汚染 激減 付いたまま 病気等 赤潮 話題ばかり 襲撃 芋洗い状態 麻酔 ブリザード 呪い ストレス 毒)
作る (作り方 作った感想 トラブル:病気 プレッシャー)
味わう (味わい方 味わった感想)
準備 (Preparation)
こだわる (こだわり方 こだわった感想)
与える (与え方 与えた感想 トラブル:プレッシャー)
つかむ (つかみ方 つかんだ感想 トラブル:プレッシャー)
育てる (育て方 育てた感想 トラブル:病気)

アオブダイ 食べ方

- 商用検索エンジンを利用

アオブダイ　毒

アオブダイ

利用 (Use)

食べる (食べ方　食べた感想　トラブル:臭い　病気　腐敗　汚染　ストレス　食物アレルギー　プレッシャー　アレルギー　ヘルペス　毒)
見る (見方　見た感想　トラブル:乱獲
ばかり　襲撃　芋洗い状態　麻酔　ブリザ
作る (作り方　作った感想　トラブル
味わう (味わい方　味わった感想)

ほほー、というわけで「毒」

準備 (Preparation)

こだわる (こだわり方　こだわった感想)
与える (与え方　与えた感想　トラブル:プレッシャー)
つかむ (つかみ方　つかんだ感想　トラブル:プレッシャー)
育てる (育て方　育てた感想　トラブル:病気)

アオブダイ　毒


ウェブ | 登録サイト | 画像 | 音声 | 動画 | ニュース | ブログ | 辞書 | 知恵袋 | エリア | 商品
YAHOO! JAPAN 検索
"アオブダイ" "毒"
検索
検索オプ
ウェブ検索結果 (検索結果の見方)
"アオブダイ" "毒" で検索した結果 1〜10件目 / 約350件 - 0.
1. 食中毒原因物質 魚の毒
魚の毒による食中毒. フグとテトロドトキシン アオブダイとパリトキシン ... 熱帯のサンゴ礁の周囲にいるシガテラ毒魚など、さまざまな形で毒を持つ魚がいます。 ... 最近ではアオブダイの切り身による食中毒事例が、１９９７年１０月に報告されています。 ...
www.ikagaku.co.jp/bac/sakana.html - キャッシュ - 5k - このサイト内で検索
PR
直撃!山本梓
ドSって
言われるけど…
Charger 3月号
[月刊チャージャー]

アオブダイ

アオブダイ

利用 (Use)

食べる (食べ方　食べた感想　トラブル:臭い　病気　腐敗　汚染　ストレス　食物アレルギー　プレッシャー　アレルギー　ヘルペス　毒)
見る (見方　見た感想　トラブル:乱獲　病気　ダブルパンチ　汚染　激減　付いたまま　病気等　赤潮　話題ばかり　襲撃　芋洗い状態　麻酔　ブリザード　呪い　ストレス　毒)
作る (作り方　作った感想　トラブル:病気　プレッシャー)
味わう (味わい方　味わった感想)

準備 (Preparation)

こだわる (こだわり方　こだわった感想)
与える (与え方　与えた感想　トラブル:プレッシャー)
つかむ (つかみ方　つかんだ感想　トラブル:プレッシャー)
育てる (育て方　育てた感想　トラブル:病気)

スクロールダウンすると

準備 (Preparation)

こだわる (こだわり方　こだわった感想)
与える (与え方　与えた感想　トラブル:プレッシャー)
つかむ (つかみ方　つかんだ感想　トラブル:プレッシャー)
育てる (育て方　育てた感想　トラブル:病気)
運ぶ (運び方　運んだ感想)
求める (求め方　求めた感想)
作る (作り方　作った感想　トラブル:病気　プレッシャー)
飼う (飼い方　飼った感想　トラブル:病気　アレルギー)
飼育 (飼育方法　飼育した感想)
集める (集め方　集めた感想)
出す (出し方　出した感想　トラブル:病気　ストレス　プレッシャー)
焼く (焼き方　焼いた感想
紹介 (紹介方法
切る (切り
選ぶ (選び
探す (探し方
調理 (調理方法
漬ける (漬
釣る (釣り方　釣った感想　トラブル:乱獲　酸欠　汚染　赤潮　プレッシャー)
買う (買い方　買った感想　トラブル:本ばかり　高騰)
売る (売り方　売った感想　トラブル:激減　風評被害　本ばかり)
販売 (販売方法　販売した感想)
料理 (料理方法　料理した感想)

もっとましな「釣り方」はあるか？

アオブダイ　釣り方

日	場所	釣果	釣り人	備考
３月２４日	南部堺・大島　平島の裏	チヌ６匹・アイ	細	

日記　細

釣り方 紀州釣り　竿：TOURNAMENT グレ １．２−５３II ﾘｰﾙ：小型ﾘｰﾙ 道糸３号　ハリス：２．０号 ハリ：がまかつ 口太グレ ５号

エサ マキエ：ヌカ ４升・砂 ２．０升・ｸﾞﾚﾊﾞﾜｰ ６５０ｇ・細引きｻﾅｷﾞ 850ｇ・ｵｼﾑｷﾞ ５００ｇ・ｱﾐｴﾋﾞ ２．５ｋｇ　サシエ：ｵｷｱﾐﾅﾏ

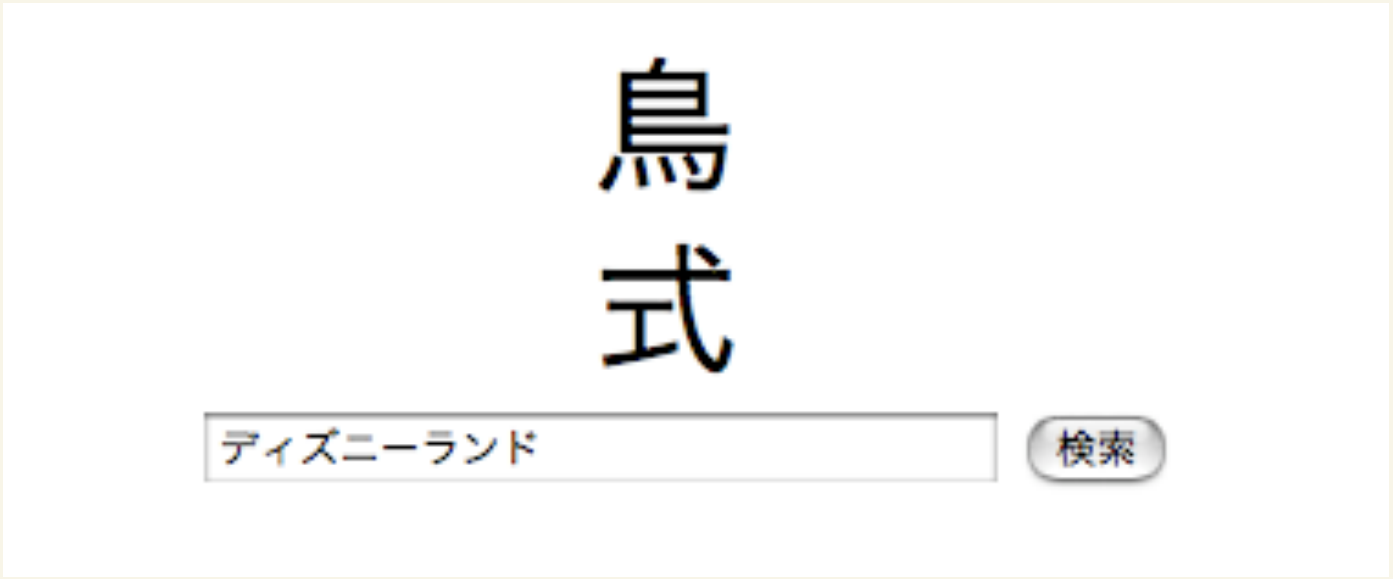
釣果 ３月２４日 午前６：００堺大島上陸。　気温 １０℃ 水温 １７．２℃
７：４５ それまでアタリは無かったのに、浮きに微妙なアタリ ４７ｃｍのチヌ ゲットこれで一安心。
７：５０浮きの変化に合わせてみると、３０ｃｍのアイゴゲット。
その後９：００まで外道のオンパレード 後アタリは止まった。
南風が強くなり釣りずらくなったので、２：００に平島のウラへ磯替りをしました。
平島のウラは背中からの風で天国であった。
しばらくしてボラ が掛かった、３：００に５１ｃｍのチヌ ゲットから入れ食い状態になり釣りまくりました。
大きさ、引数ともに自己記録更新ができたことがなによりの一日であった。

釣果は、　チヌ ４２ｃｍ ４３ｃｍ （４５ｃｍ） ２枚 ４７ｃｍ ５１ｃｍ

他魚は、ボラ １匹、アオブダイ２匹

ＴＯＰへ　釣り日記へ

ディズニーランド

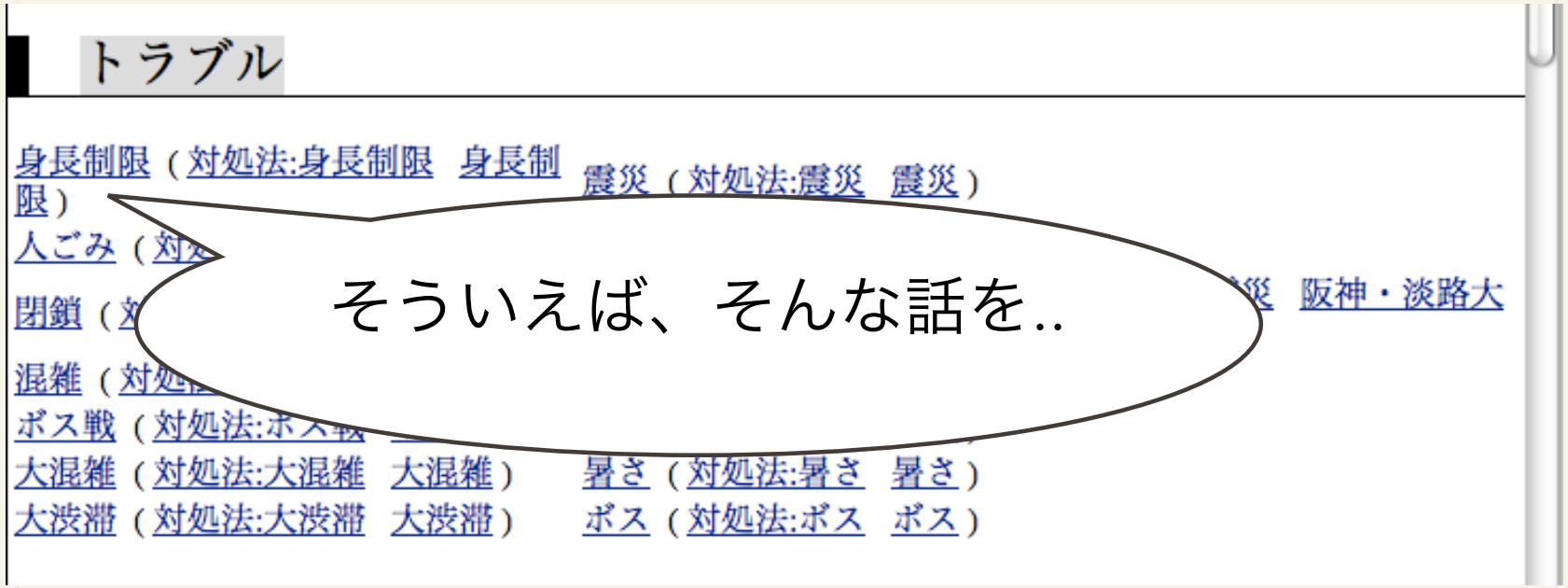


ディズニーランド


ディズニーランド
利用 (Use)
観る (観方　観た感想)
読む (読み方　読んだ感想)
見る (見方　見た感想)
知る (知り方　知った感想)
見つける (見つけ方　見つけた感想)
歩く (歩き方　歩いた感想)
過ごす (過ごし方　過ごした感想)
遊ぶ (遊び方　遊んだ感想)

ディズニーランド

- スクロールダウンすると



ディズニーランド

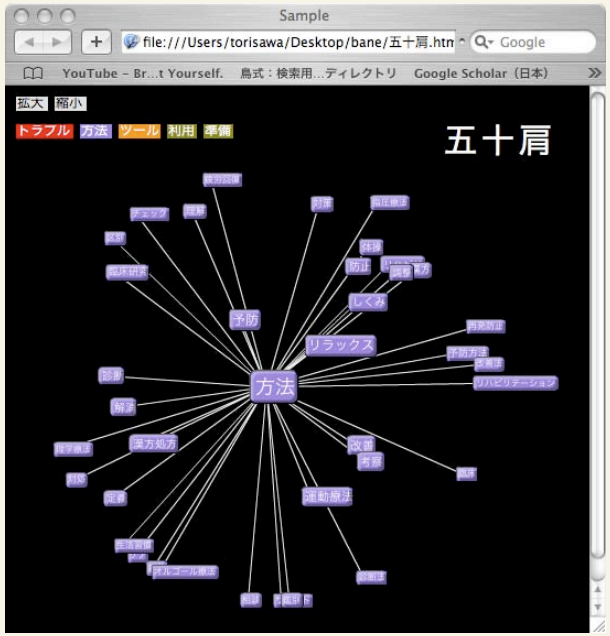
- 身長制限をクリックすると

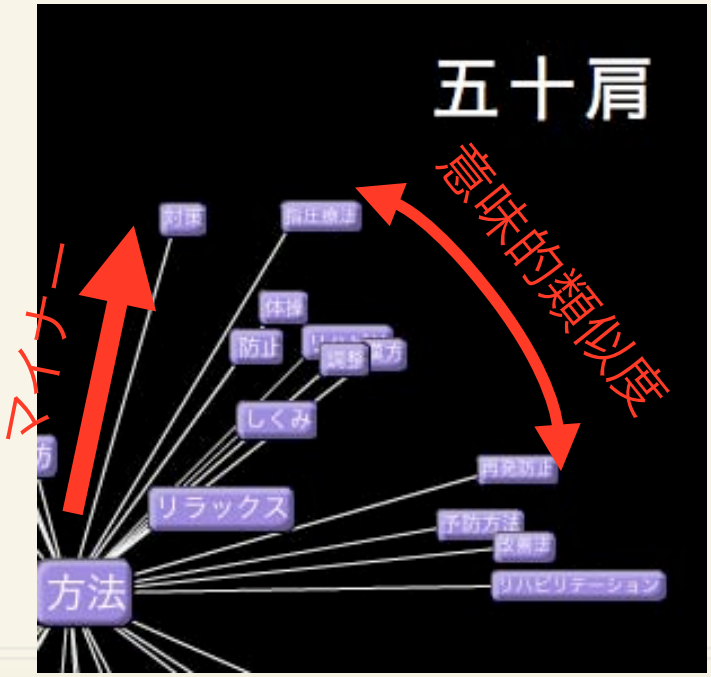

1. ディズニーランド
... 注意するアトラクション ・注意するアトラクション (妊婦編) ・注意するアトラクション (幼児編) ・年齢・**身長制限**のあるアトラクション ・東京にしかないアトラクション ・雨の日も安心アトラクション& ショー・パレード ・ショー ...
homepage2.nifty.com/smee/land.html - キャッシュ - 101k
- 2007年2月20日 - このサイト内で検索

- 身長制限のあるアトラクション一覧へのリンク有り
- 公式サイトで探すのは結構大変

デモ

- 通常のブラウザーで動作可能(Flashを利用)





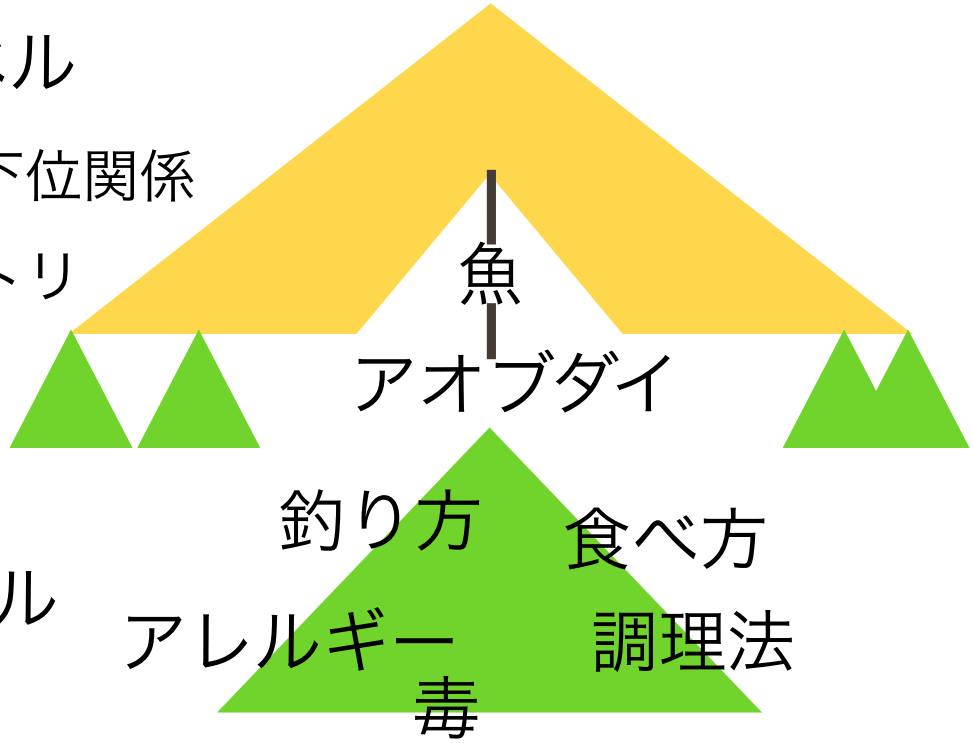
構造

- 二階層

対象特定レベル

（隅田, 2007）上位下位関係

現状120万エントリ



様相特定レベル

釣り方　食べ方

アレルギー　調理法

毒

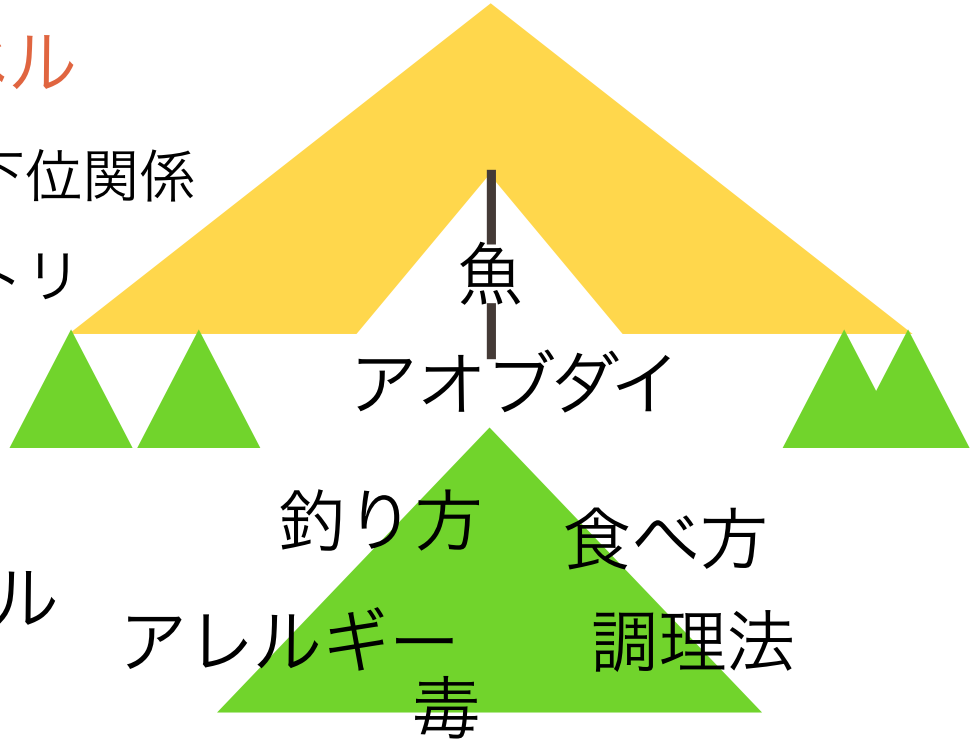
構造

- 二階層

対象特定レベル

（隅田, 2007）上位下位関係

現状120万エントリ



対象特定レベル

- シソーラス
  - Webコーパス（テキストデータ）から「上位下位関係」を自動抽出
    - 「アオブダイなどの魚」
    - HTML文書中のレイアウト＆統計量
    - Wikipediaから抽出
  - 現在、精度約８０％

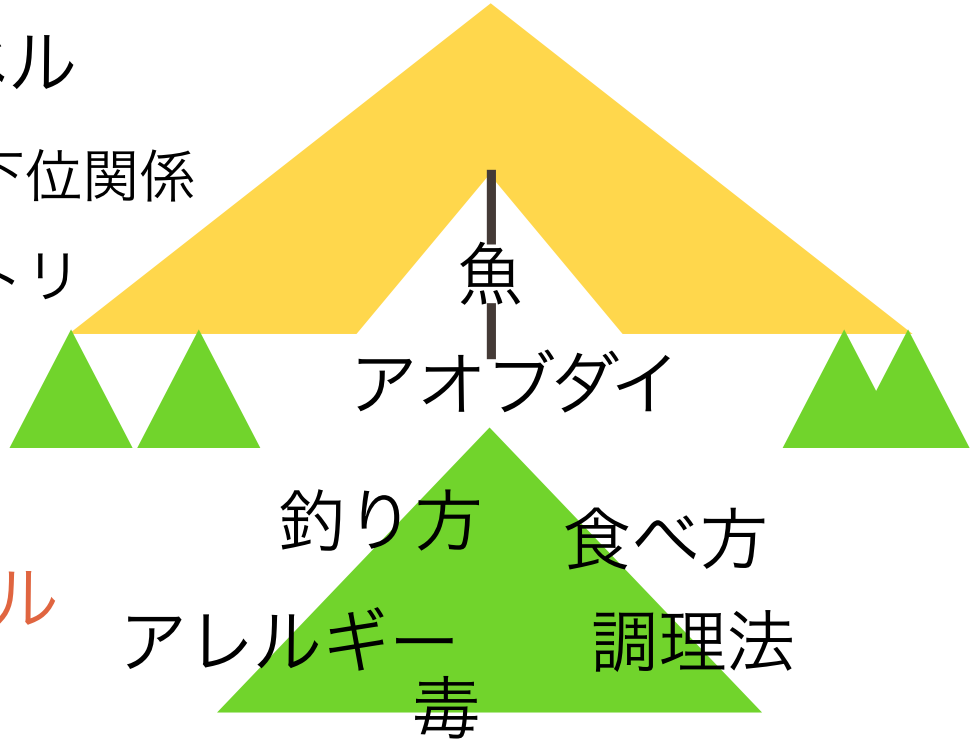
構造

- 二階層

対象特定レベル

（隅田, 2007）上位下位関係

現状120万エントリ



様相特定レベル

様相特定レベル

- 対象の利用（e.g., 魚→食べる）
  - 利用の方法、コメント、トラブル、道具、具体的方法、...
- 利用の準備 （e.g., 魚→釣る、買う、飼う）
  - 準備の方法、コメント、トラブル、道具、具体的方法、...

シナリオ：利用、準備を表す動詞

- e.g, アオブダイ
  - 利用：食べる、準備：釣る、買う
- 獲得手続（鳥澤,2005）
  - 高頻度で検索対象と共起する動詞
  - 助詞の「で」にバイアス
    - e.g., 「ディズニーランドで遊ぶ」
  - 人手で作成した学習データを用いた教師あり学習
    - ある名詞の利用を表す動詞は、他の名詞の利用を表す動詞にもなりやすい

トラブル表現

- 手がかり
  - 係り受け関係（動詞の否定／肯定を区別）

・否定形の動詞との係り受け関係が重要な手がかり

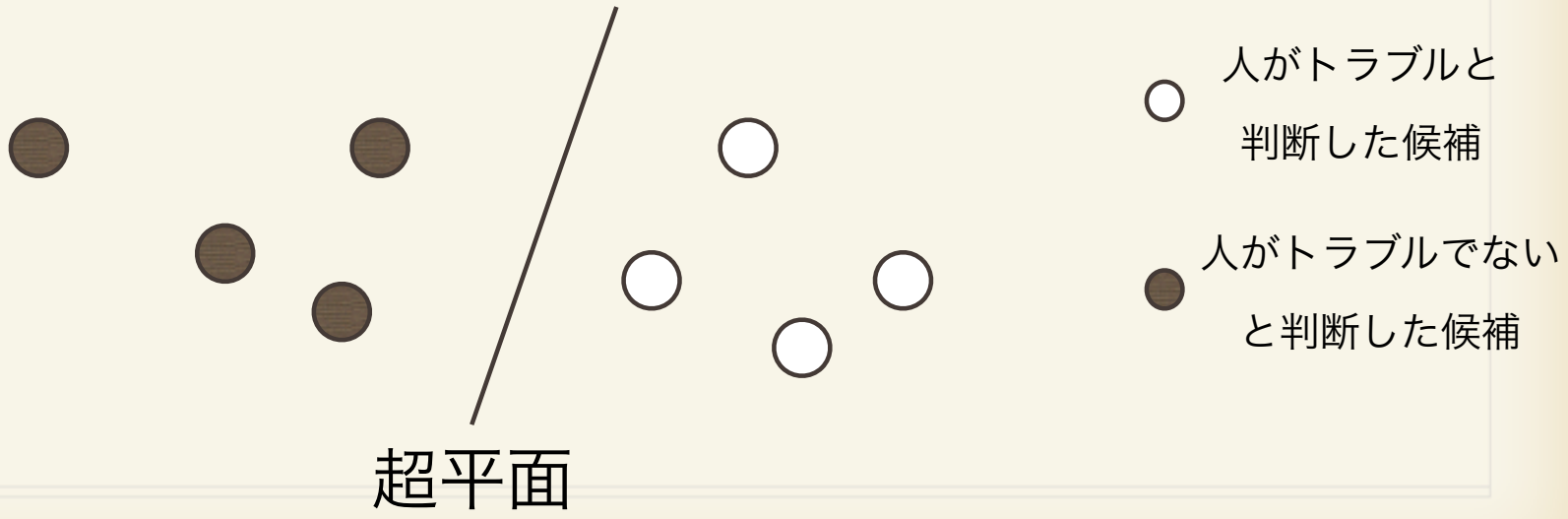
　・トラブルとは何か行為を不可能にするもの

・ただし肯定形が手がかりになることもある

　・e.g.,「困る」「死ぬ」「苦しむ」

トラブル表現

- 機械学習で複数の証拠を組み合せる
  - 機械学習（SVM）：トラブル／非トラブルを分離する超平面を計算

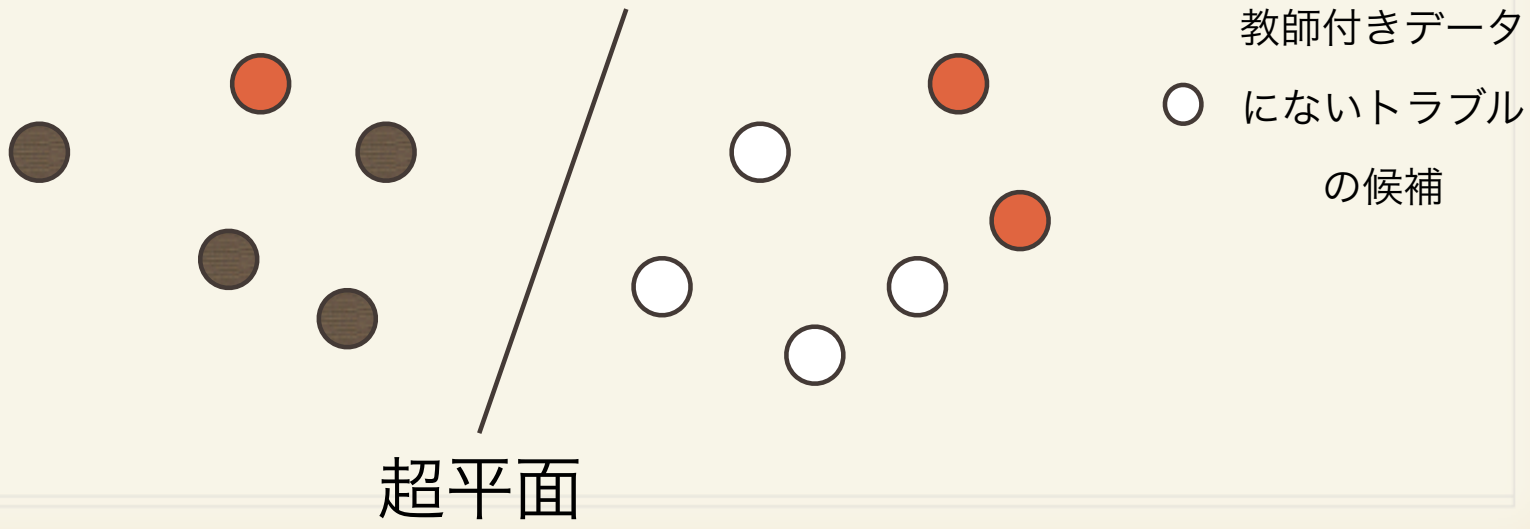


トラブル表現

- 機械学習で複数の証拠を組み合せる
  - 素性ベクトル：トラブルの候補各々に一つ
    - 教師付きデータ：少数のサンプルを人手でトラブルかどうか判断しておく

トラブル表現

- 機械学習で複数の証拠を組み合せる
  - 素性ベクトル：トラブルの候補各々に一つ
    - 教師付きデータにない候補も超平面から判断

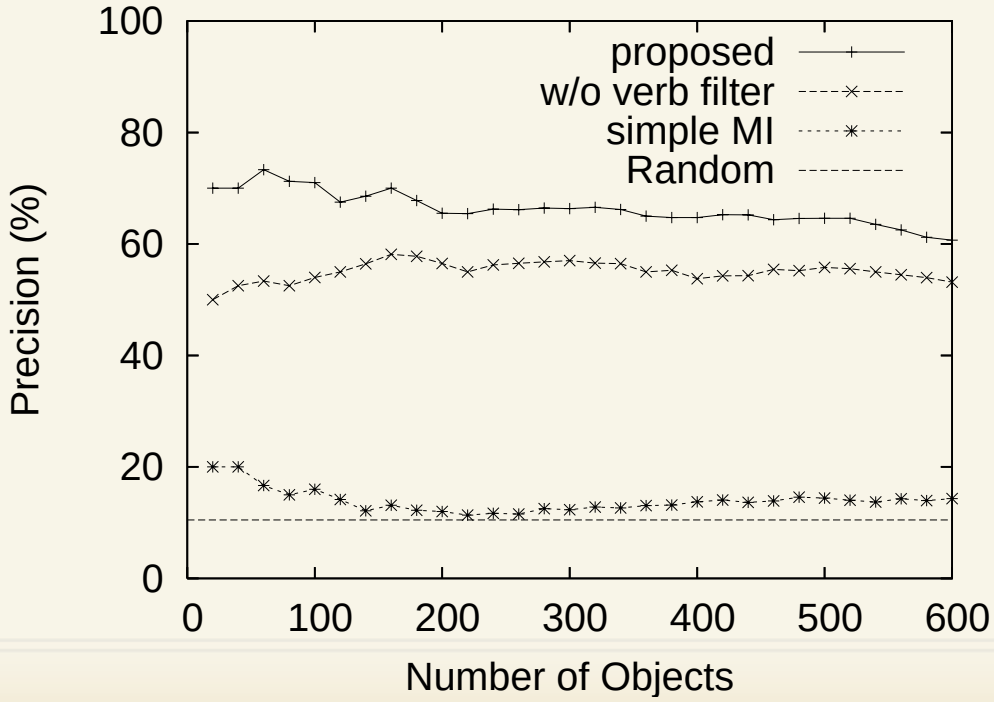


トラブル表現

- 想定外な出力例
  - 鯖落ち、バグバグ、五十肩、雨漏り、ブルー画面、人人人、定休日、未実装、寝癖

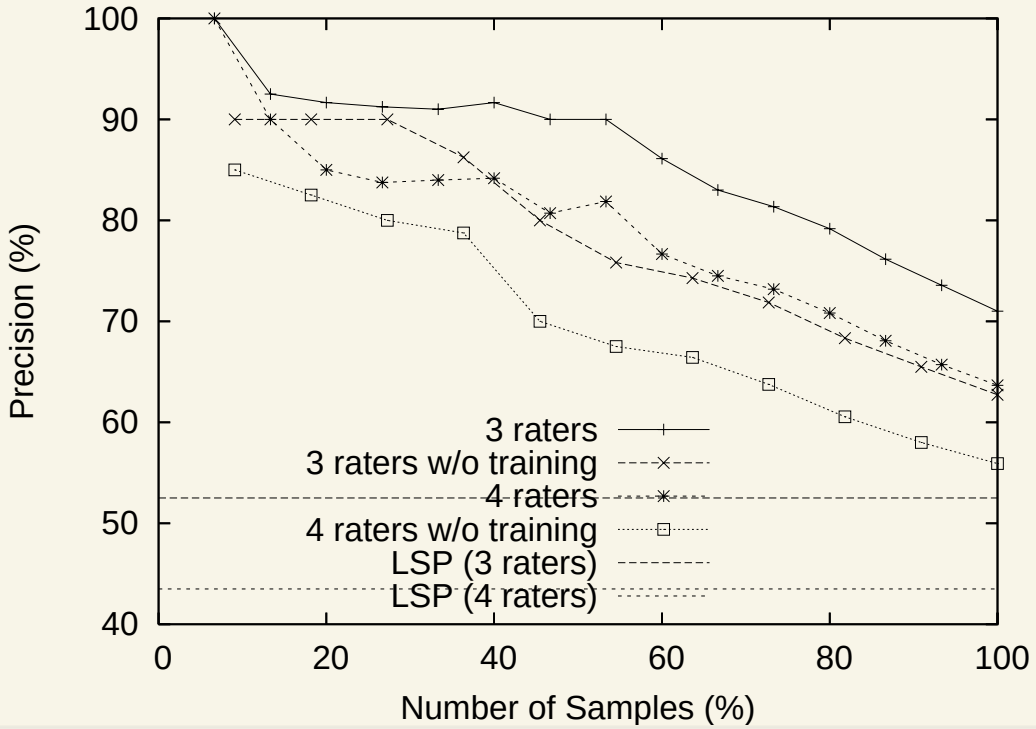
トラブル表現

- 検索対象とトラブル表現の対応関係の制度（＜遊園地、身長制限＞）
  - ４名の被験者中３名がOK



トラブル表現

- トラブル表現の精度（全部で3000個）
- 出力結果を４人の被験者が判定



トラブル以外

- 合計9万個の正解付きサンプルを作成、
  - Active Learning, ヒューリスティックス

	Samples (Positive Samples)	Precision	Acquired Num (Training Samples)
方法	30,000(9,312)	0.74	137,756(9,312)
道具	30,000(11,992)	0.815	119,778(11,992)
場所	30,000 (11,858)	0.815	73,160(11,858)

トラブル表現

- 対象と「その利用で予想されるトラブル」の対を求める
  - 相互情報量による関連づけ
    - e.g., 遊園地→身長制限
      - 「遊園地の身長制限」は、「遊園地」「身長制限」それぞれの出現確率から予想されるものよりどれだけ多く出現しているか？

トラブル表現

- どの利用法、どの準備法に関係したトラブルか？
  - e.g., 遊園地→遊ぶ→トラブル：身長制限
  - 係り受けの利用
    - e.g., 「身長制限で遊べない」、「遊園地で遊ぶ」の二つの係り受

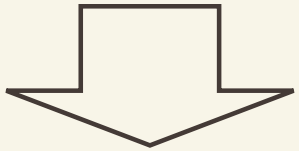
      ⇒身長制限、遊ぶ、遊園地の三者に関係ありと認識

まとめ

- 対象の利用のシナリオに沿って、つまり対象の利用価値（Heidegger,Marx,...)に沿って、価値あるアイテムを提示
  - 様々な価値が利用価値から派生
    - 毒、事故⇒安心安全
    - 「鮎」を「買う」⇒経済的最適化
    - 「鮎」の「ルアー釣り」⇒イノベーション

まとめ

- ネット上から興味／フォーカスにそった価値ある想定外の発見を支援



  - つぼを得た知識流通支援へ

- キーとなる技術：自然言語処理による大量のWeb文書からの知識獲得
  - 価値に関するビジョン毎に異なる知識獲得技術が必要

情報爆発成果報告会

日　時：平成２０年３月３日(月)～４日(火)

場　所：秋葉原コンベンションホール

東京都千代田区外神田１-１８-１３

秋葉原ダイビル２F、５F　(JR秋葉原駅から徒歩１分)

http://www.akibahall.jp/data/access.html

主　催：文部科学省科学研究費補助金 特定領域研究

「情報爆発時代に向けた新しいIT基盤技術の研究」

http://www.infoplosion.nii.ac.jp/info-plosion/

後　援：国立情報学研究所(予定)

参加費：無料（事前に参加登録をお願い致します）

成果報告会・参加登録URL：

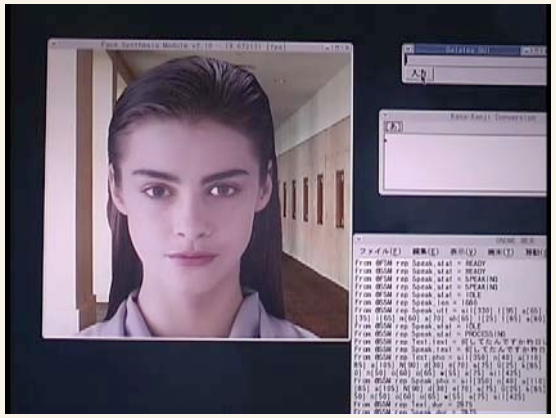
http://www.infoplosion.nii.ac.jp/info-plosion/html/houkokukai-h19/

今後の予定

- 価値を反映した言語資源の拡充、精度向上
  - さらなる機械学習の導入
- サーチの改善
  - 現在は単なるboolean search
  - TSUBAKI（京大黒橋研）の利用
  - 自動獲得された「言い換え」の導入
    - 「5.1ch再生する」⇒「聞く」

他のアプリケーション

- 対話エージェントによる有用な情報の一般ユーザーからの収集、ユーザー間のコミュニッケーションの仲介



鳥澤研HPからデモ視聴可能

- 自動獲得した知識をもとに、多様なトピックの利用価値に関係した質問を自動生成、入力理解は最小限