# 揺らぎのある言葉と揺らぎのない情報検索

慶應義塾大学 理工学部情報工学科/デジタルメディア・コンテンツ統合研究センター　金子 晋丈

Oxford Day 2018/10/21 @慶應義塾大学

## Overview

- 自己紹介
- 背景 - データの利活用が進まない -
- 検索のメカニズム
- 検索サイトの構築
- 慶應DMCのアプローチ



## 自己紹介

- 金子が最近やってること：「コンテンツの長期的な利活用」
  - コンテンツネットワークが鍵
    - デジタルデータをDLするためのネットワーク(コンピュータNW)から
      デジタルデータに出会うためのネットワーク(コンテンツNW)へ
  - JIS「Z 6019 磁気テープによるデジタル情報の長期保存方法」原案作成
- 映画・映像関係
  - DCI Compliance Testの実施 (映画館機材の性能・セキュリティ検証)
  - CineGrid: 4Kコンテンツの流通



## データの利活用が進まない

- ITの現状
  - ブロードバンド、十分なストレージ、コンピュータ
  - 情報発見は、Google, Facebook, テレビ, 意外と昔ながらの人づて？
- 消費者の見たいコンテンツとは？
  - 全く知らない情報 (見向きもされない場合あり)
  - あまり詳しくないが、興味のもてる情報
  - 詳しい知識をさらに深める情報

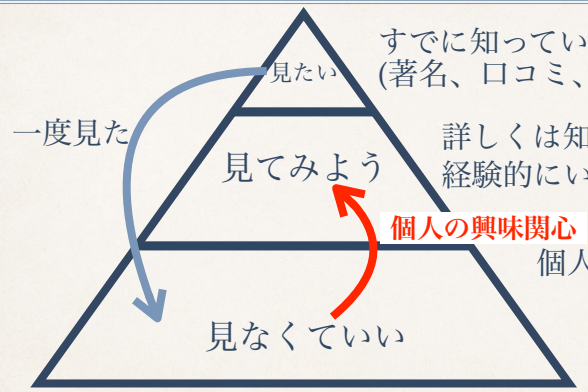
**どの手法がどのコンテンツの閲覧に貢献しているのか？**

# 見たいコンテンツとは？

- コンテンツの類型
  - **見たいコンテンツ：Amazon**
    - すでにタイトル等を知っていて、見たいと思っている
    - 検索してコンテンツを発見し視聴 (すでに売り上げに貢献)
    - 例：Paris に行ったらルーブル美術館、エッフェル塔
  - **見てみようと思うコンテンツ：書店**
    - 見たことないし、聞いたこともないが、見てみたいと思う
    - 一度見たけど、改めてもう一度見てみたいと思う
    - コンテクスト（その時の雰囲気や環境）が閲覧に（売り上げ未貢献）
    - 例：Paris の街歩きでふらっと入るカフェ
  - **見なくていいと思うコンテンツ：自宅の書棚**
    - 一度見たコンテンツ、全く興味のないコンテンツ



# コンテンツピラミッド




すでに知っている=検索可
(著名、口コミ、広告・宣伝)

詳しくは知らない(検索不可)が、
経験的にいいかも、と思える

個人に馴染む形(コンテクスト)で
コンテンツへの
接触面を増やす

コンテンツネットワークの研究：ネットワーク的アプローチでこの問題に挑む



# 検索のメカニズム

# 検索実現の基本メカニズム 1/2

- 言語表現
  - file_red.txt: This is a red pen.
  - file_blue.txt: This is a blue pen.
- コンピュータ表現
  - T h i s i s a r e d p e n .
    0x54 0x68 0x69 0x73 0x20 0x69 0x73 0x20 0x61 0x20 0x72 0x65 0x64 0x20 0x70 0x65 0x6e 0x2e
  - T h i s i s a b l u e p e n .
    0x54 0x68 0x69 0x73 0x20 0x69 0x73 0x20 0x61 0x20 0x62 0x6c 0x75 0x65 0x20 0x70 0x65 0x6e 0x2e
- スペース(0x20)やピリオド(0x2e)を見つけて、各単語の取り出し

# 検索実現の基本メカニズム 2/2

- file_red.txt
  - 0x54 0x68 0x69 0x73 0x20 0x69 0x73 0x20 0x61 0x20 0x72 0x65 0x64 0x20 0x70 0x65 0x6e 0x2e
- スペース(0x20)やピリオド(0x2e)を見つけて、各単語の取り出し

  T h i s i s a r e d p e n .
- 各単語を含むファイル名の表を作成
  - 0x54 0x68 0x69 0x73 (This): file_red.txt, file_blue.txt
    0x69 0x73 (is): file_red.txt, file_blue.txt
    0x72 0x65 0x64 (red): file_red.txt
    0x62 0x6c 0x75 0x65 (blue): file_blue.txt
    0x70 0x65 0x6e (pen): file_red.txt, file_blue.txt

検索キーワード

0x72 0x65 0x64 ← red

0x52 0x45 0x44 ← RED

0x52 0x65 0x64 ← Red



# 検索キーワードの拡張

- 検索キーワードの文字列が完全マッチしないとヒットしない
  - 大文字、小文字が違っていてもヒットして欲しい
- 検索キーワードを「辞書」に従って生成
  - 利用者の入力: 0x52 0x45 0x44 (RED)
- 検索エンジンによる拡張: 0x52 0x45 0x44 (RED), 0x72 0x45 0x44 (rED), 0x52 0x65 0x44 (ReD), 0x52 0x45 0x64 (REd), 0x72 0x65 0x44 (reD), 0x72 0x45 0x64 (rEd), 0x52 0x65 0x64 (Red), 0x72 0x65 0x64 (red)
- 「辞書」
  - 0x52=0x72 (R = r), 0x45=0x65 (E = e), 0x44=0x64 (D = d)

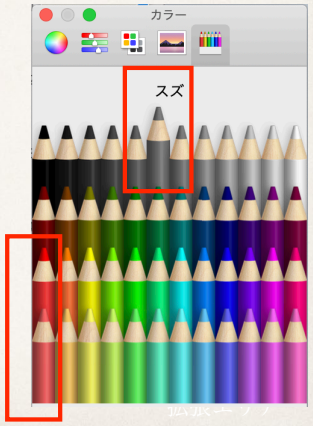


# 「辞書」の拡張

- 「辞書」は一種の類義語辞典
- 辞書の拡張
  - 赤 = 朱 = さくらんぼ = サーモン
  - グレー = スチール = スズ = ニッケル

    **??**
  - 色としてみたら類義語
  - しかし、常には類義語になり得ない

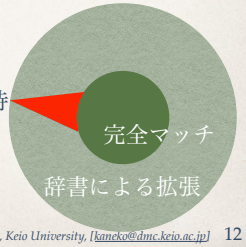
**汎用的な類義語辞典は存在しない**






# コンテクストが失われた類義語辞典

- 検索キーワードの文字列が完全マッチしないとヒットしない
  - 大文字、小文字が違ってもヒットして欲しい
  - スペルが間違っていてもヒットして欲しい
  - 同じような意味を持つ単語であればヒットして欲しい
  - 他言語でも同じような意味を持つ単語であればヒットして欲しい
  - …

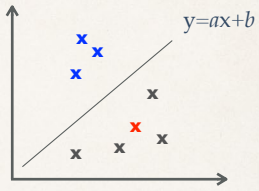



**辞書が検索者の期待と一致しているか不明**

(検索エンジンには検索者の期待を把握するメカニズムがない)

# 人工知能？

- 人工知能(AI)
  - 教師データに基づいて処理基準を汎用化
  - 教師データに最適化した画一基準
  - すべての入力データに完璧に対応できない
- 計算機で自然言語に挑む
  - 完全一致を求める計算機と意味が揺らぐ言葉とのギャップ
  - **揺らぎをどのように画一的な計算で吸収するのか？**
    - 現行手法の限界：語彙統制、辞書の導入、人工知能の導入
    - 画一的な手法を**どこで**動かすと一番柔軟な処理ができるのか？






# 検索サイトの構築

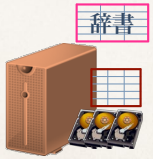
# 検索サイト

- 検索サイト
  - キーワードindex(キーワードとファイル名の対応付け)を保持
  - 「辞書」を保持
- なるべく大きなindexを保持することが重要
  - 利用者は一回の検索でできるだけ多くの情報を集めたい
- 検索サイトの運営主体による違い



# 情報保有者の直接提供モデル

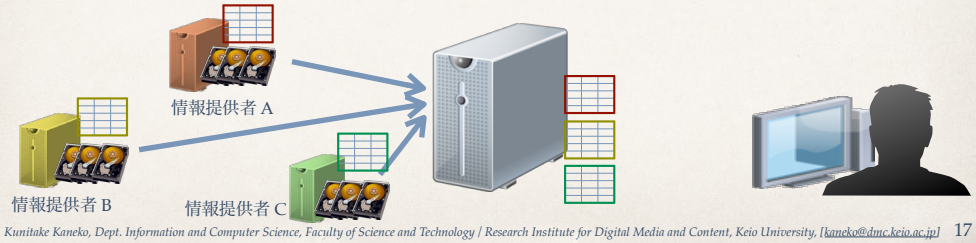
- 提供者自身による情報検索の提供
  - 提供できる情報は限定的
  - 利用者に存在を知らせる方法がない
  - 情報の範囲と利用者が限定的なので、
    コンテクストを明確化した辞書作りができる




情報提供者
A

# 第三者による間接提供モデル

- さまざまな情報を一括検索できる
- 多くのサイトの情報を集めなければ、検索サービスを提供できない
  - 情報提供してくれるか？知財、知名度…
- 情報の範囲と利用者が広範なので、コンテクストが不明瞭
  - 辞書の導入が過剰な検索ヒット数を招き、利便性低下





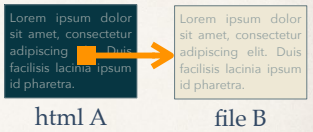

# どこで検索サイトを運営するか？

- 検索サイトの運営場所は一長一短
  - 提供者自身による検索サイト
  - 第三者による検索サイト
- **利用者自身による検索サイトはできないのか？**
  - 情報提供者を見つけることができれば可能
  - 利用者のコンテクストを取得可能
    - 自分だけの特別辞書を導入



# webにおいて情報提供者を見つけることは可能か？

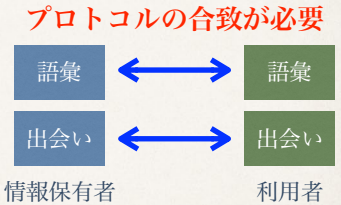
- WEBの構造
  - HTMLのリンクによるグラフ構造
  - あるhtmlからの外向きのlinkしか記述できない
    - AからBを見つけられるが、BからAは見つけられない
- 情報提供者の存在を知るには、web空間の全探索が必要
  - web空間の全ファイルの取得には膨大なコンピュータが必要
  - google等の大手以外は参入できない実態
- **グラフの作り方を変えると情報提供者を容易に発見可能**
  - 例：blogのtrackback






# 検索技術のまとめ

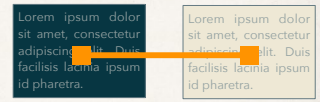
- 検索により情報を提示できるための2つの条件
  - 情報保有者と利用者が同じ語彙を持つ
  - 情報保有者と利用者の出会いの場が存在する
- 検索技術の高機能化
  - 語彙の補完を行う
  - 多くの情報保有者を呼び寄せる
- 出会いの場の課題
  - 語彙が合致しないために出会えない
  - 語彙が合致しすぎるために出会えない



# 慶應DMCのアプローチ

## DMCのアプローチ: Catalogue System

- 関係するデータを利用者が直接発見できるように
  - AからBをBからAを発見可能
- なるべく多くの情報提供者を見つけられるように
  - fileを保有していなくてもリンク可能
- データ間を言語化せずに関係づけられるように
  - 言語化できない数値情報等も取り扱い可能
- 利用者側で柔軟性が必要な処理を実行できるように
  - 言葉を用いた検索等、パーソナライズ

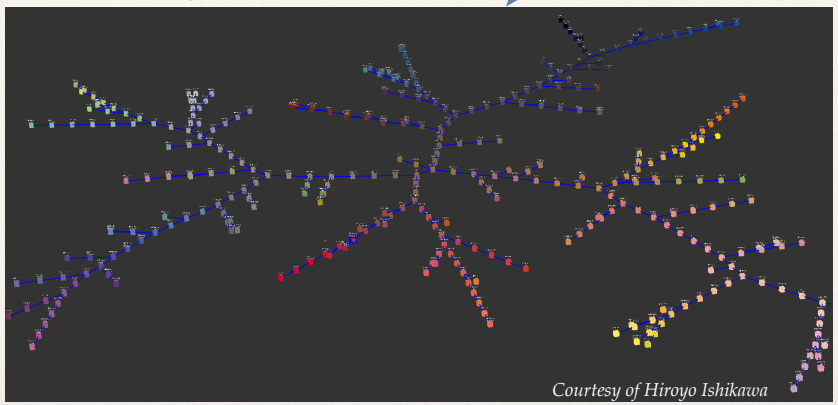





## Catalogueでつなげた色の世界

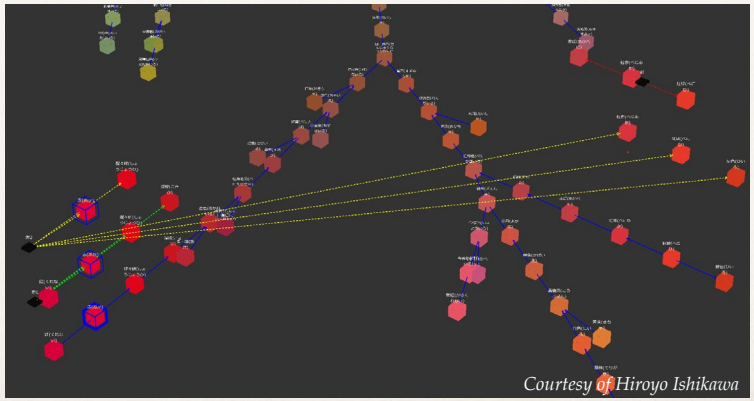
- MoSaICによる可視化

言語で色の関連性を記述できますか？



# Catalogueでつなげた色の世界

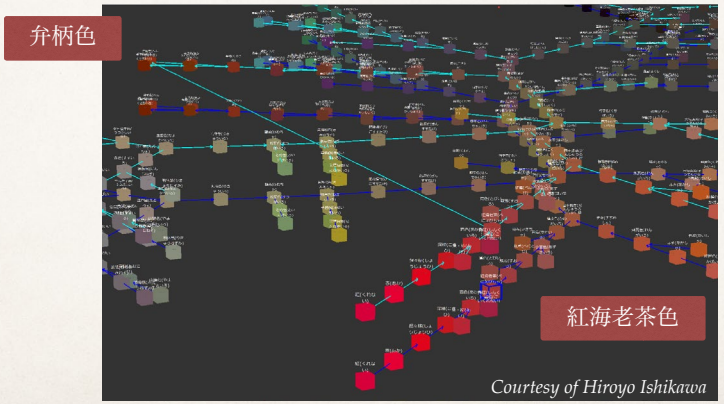
- どこからどこまで赤色？朱色？






# Catalogueでつなげた色の世界

- 異なる関係づけが色を立体的につなげる。新しい色との出会い。






# パーソナライズ: 柔軟性の提供

- あるコンテンツに関係あるコンテンツだけを取得（局所取得）
  - 詳細分析・カスタマイズ解析が可能
  - 検索結果のヒット数が現実的な数に収まる
- Catalogue のグラフと個人の閲覧済みコンテンツ群/作成Catalogueの差分
  - 知ってそうで知らなかったコンテンツの提示が可能
    - 似たコンテンツ特性を持つユーザの発見
    - 似たCatalogue 特性を持つユーザの発見
  - ストーリ性のある情報の提示、ナビゲーション
- Catalogueによるコンテンツ間の紐帯強度、経路履歴、アクセス履歴
  - 利用動向、利用傾向の把握



# まとめ

- 背景 - データの利活用が進まない -
- 検索の難しさは、語彙と出会いの場
  - デジタルは画一的、自然言語(意味の捉え方)は連続的かつ柔軟
  - デジタルのための語彙統制は本末転倒
- 慶應DMCのアプローチ
  - Catalogue Systemによる画一性と柔軟性の両立