

大学図書館の活用と情報探索(第5回)

情報の組織化と検索： 理論と実際

京都大学情報学研究科

黒橋禎夫

kuro@i.kyoto-u.ac.jp



自然言語処理

(Natural Language Processing)

言葉とコンピュータ

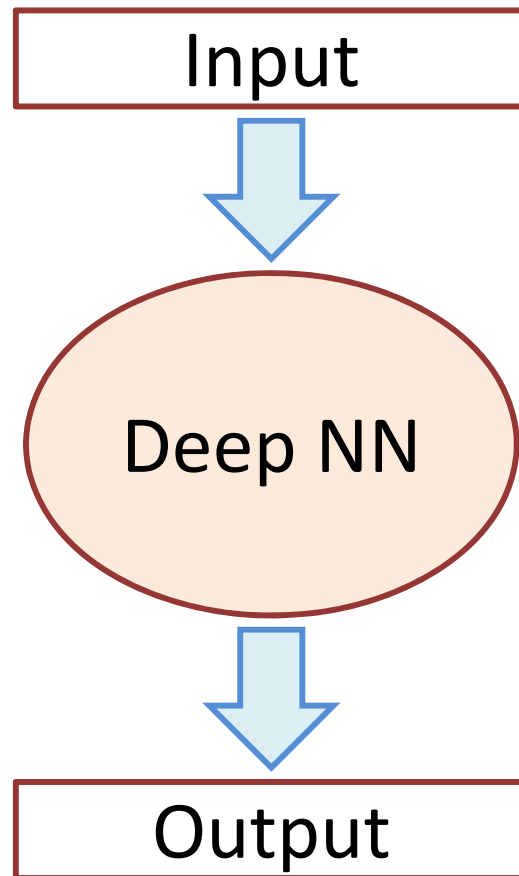


- かな漢字変換
- メール送受信
- 電子書籍
- ウェブ検索
- 機械翻訳

人間の知的な振る舞い(言語活動)を
コンピュータで実現する

- 自然言語処理 (Natural Language Processing)
- 人工知能 (Artificial Intelligence)

End-to-End Learning/System



Weather Forecast Generation from Numerical Prediction Data

Generating Video Description

[Laokulrat+, COLING2016]



A boy is playing a piano

情報の分類と組織化

- 「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)

LATCH: 5つの整理棚

(Richard Saul Wurman)

1. Location (場所)
2. Alphabet (アルファベット、50音順)
3. Time (時間)
4. Category (カテゴリ)
5. Hierarchy (序列)

分類の演習

分類の演習

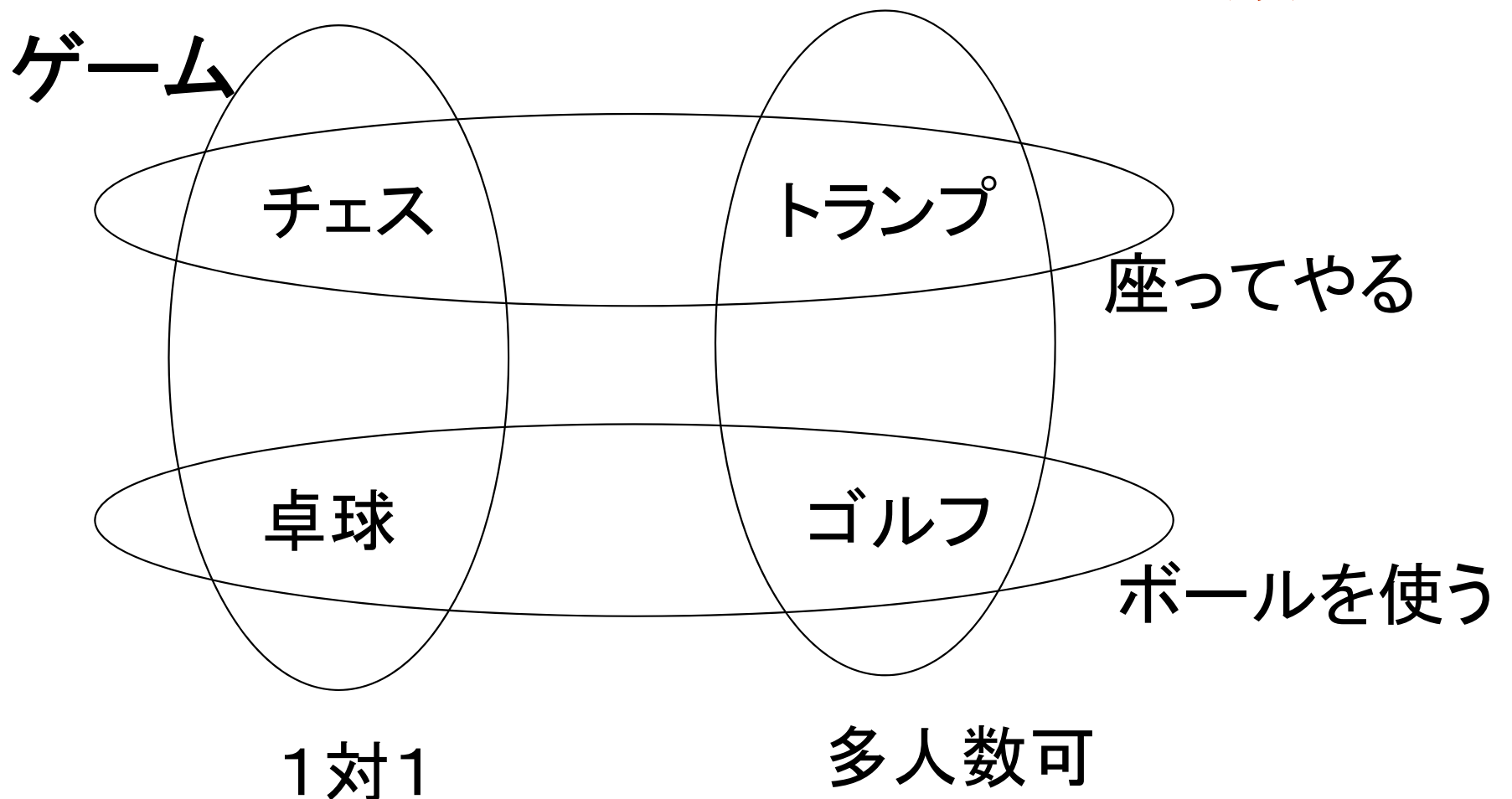
なす、新聞、ほうき、キカイダー、にわとり、
リンゴ、学生、いす、トマト、コンピュータ、
ピラニア、テレビ、掃除機、くじら

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(1889-1951)

家族的類似性



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、...

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 (Uppsala, Sweden)



- アダンプソンの植物分類
 - 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

人為分類 vs. 類型分類

- 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 → 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

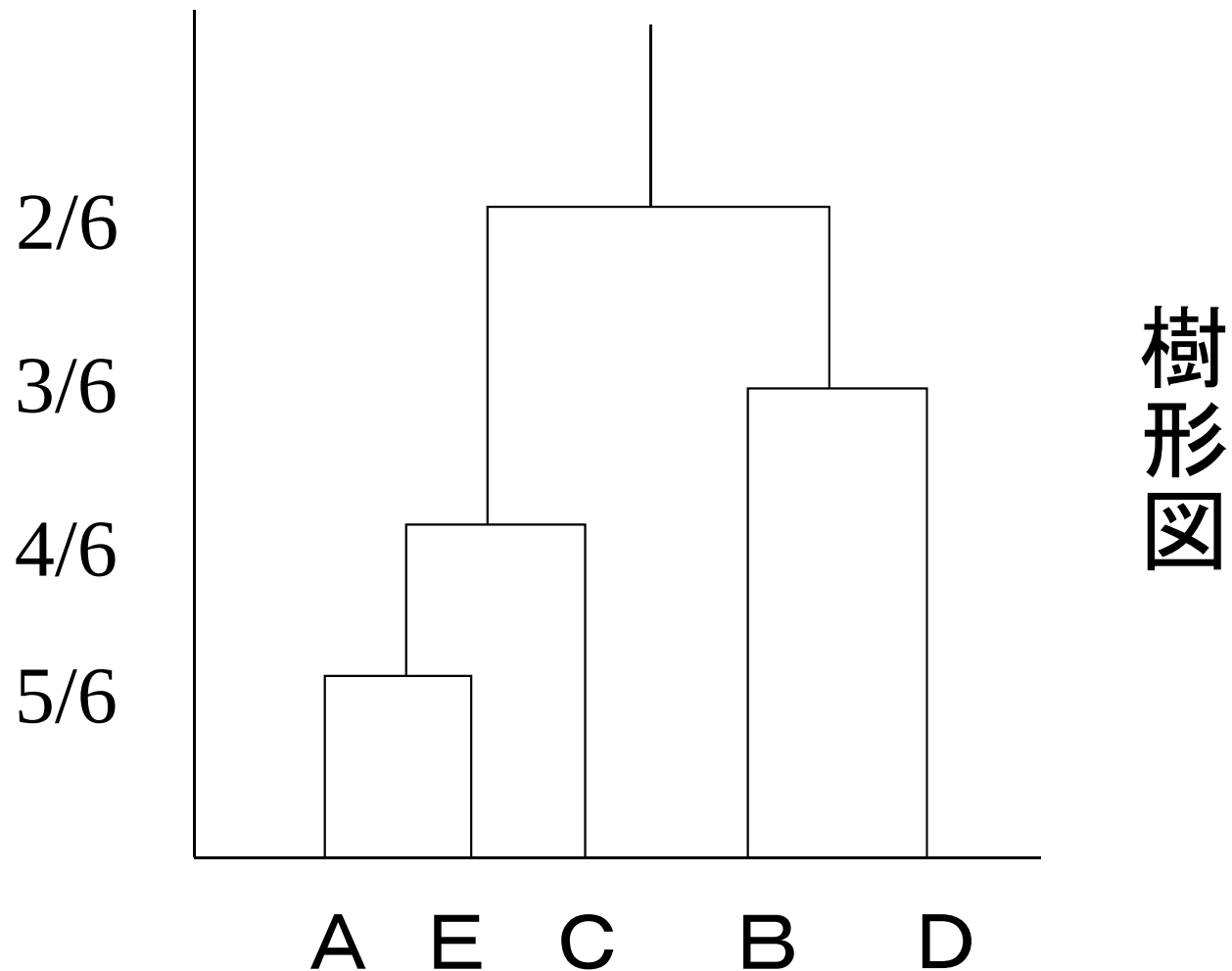
個体

	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



図書の種類

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

- 語の上位下位、同義関係などを体系的に整理したもの
 - Roget's Thesaurus (1852)
- 一般用語については同義関係が中心、単語の選択の手助け

e.g. 種々の類語辞典

分類語彙表(国立国語研)

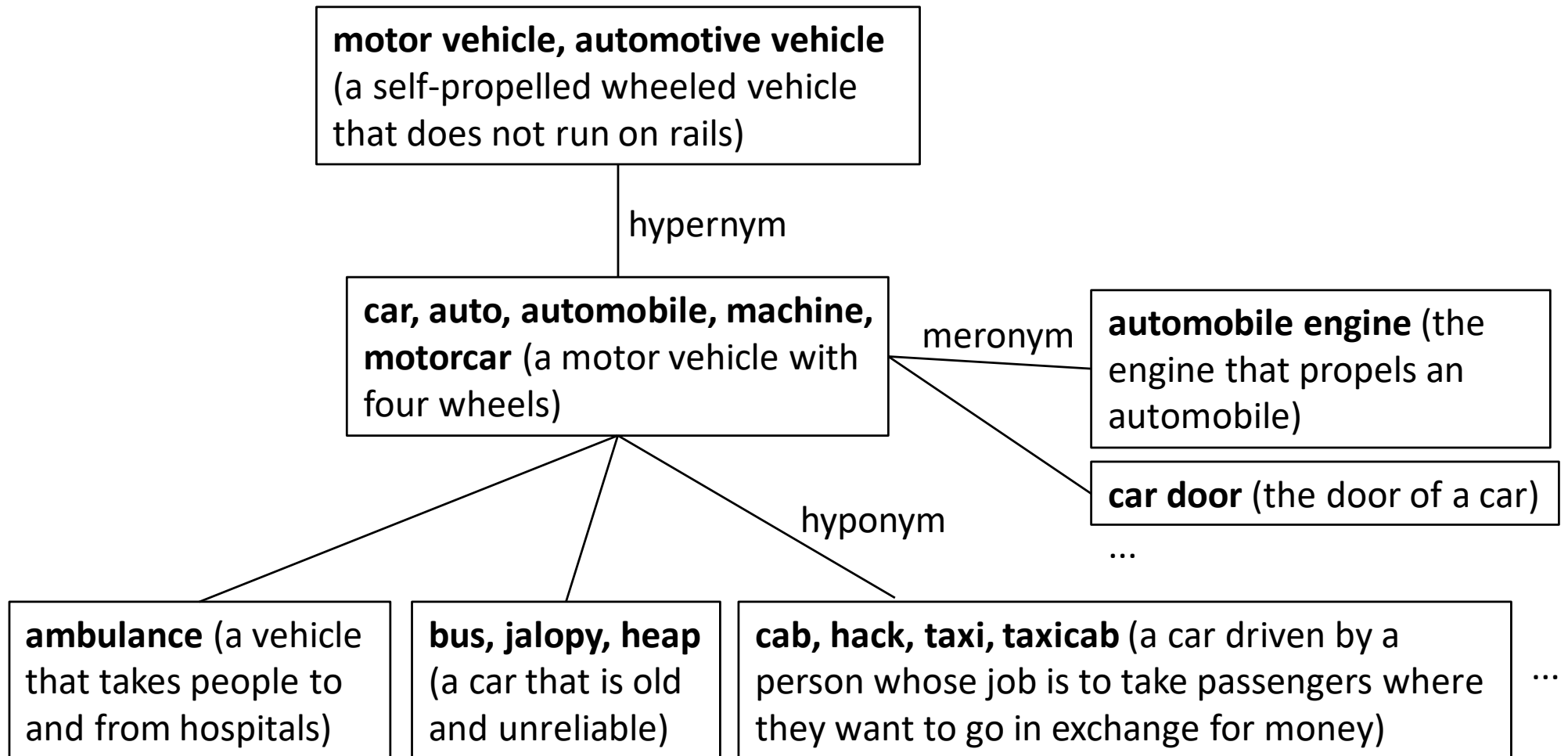
NTT日本語語彙大系

日本語使いさばき辞典

WordNet

Longman Language Activator

WordNet



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

e.g. JSTシソーラス(4万語)

情報検索

情報検索（文書検索）

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
- Webページ検索
- 社内文書検索

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



エンターテインメント

芸能人 音楽 コミック、アニメ 映画
占い テーマパーク グラビアアイドル
面白画像 お笑い 懸賞 ウェブ漫画
将棋 オススメサイト AKB48 増城
風 エヴァ ...

趣味とスポーツ

スポーツ 車 ゲーム 旅 交通
時刻表 ギャンブル アウトドア バイク
格闘技 野球 サッカー ゴルフ F1
釣り プロ野球 Jリーグ パズル
数独 ...

各種資料と情報源

辞書、辞典 図書館 郵便 郵便番号検索
電話番号と住所 カレンダー 昔の新聞
古地図 ...

芸術と人文

写真 文学 美術館 絵画 演劇 歴史
小説家 日記 建築 ダンス イラスト
古写真 昭和 廃墟 コスプレ ...

健康と医学

病気 病院 医学 ダイエット
妊娠、出産 女性の健康 癌 薬 うつ

メディアとニュース

新聞 テレビ ラジオ 雑誌 天気
番組表 動画 芸能、スポーツ新聞
2chまとめ 女子アナ テレビアニメ
ドラマ ロケ地 噂 エンタメニュース
週刊誌 ...

ビジネスと経済

ショッピング 企業間取引 不動産 求人
金融と投資 職業と雇用 企業 株
為替 テレビ局 世界のYahoo! 旅行社
倒産 ...

生活と文化

グルメ 暮らし ファッション 美容
住まい レシピ 恋愛 結婚 都市伝説
出会い ラーメン 絵日記 事故物件
ヘアスタイル 花見 GW ...

コンピュータとインターネット

インターネット ブログ ソフト 携帯
スマホ 著名人のブログ ITニュース
SNS 画像 著名人のTwitter
著名人のFacebook YouTube ...

教育

大学 資格 専門学校 予備校
小中高校 幼稚園と保育園 学校裏サイト

二国理太郎

名優が90歳で死去。

布施明

歌手の森川由加里と結婚！ 公式サイトで
報告しています。

FUNKY MONKEY BABYS

ファンキー加藤が結婚をブログで発表！

ももいろクローバーZ

西武ドームで2日間のライブが成功！ ロ
ックフェス「オズフェス」にも出演決定。

浅田真央

来年のソチ五輪後に引退の意向。

お知らせ

もっと見る

2013年04月15日

15の新しいカテゴリを追加 **NEW!**

2013年04月15日

117の新しいサイトを追加 **NEW!**

2013年03月13日

カテゴリが全面リニューアル

Yahoo!カテゴリの公式アカウント



幅広い最新情報をお届けします

いいね!

293

フォローする

Yahoo!カテゴリの最新情報を受け取る



RSS

携帯電話で見る

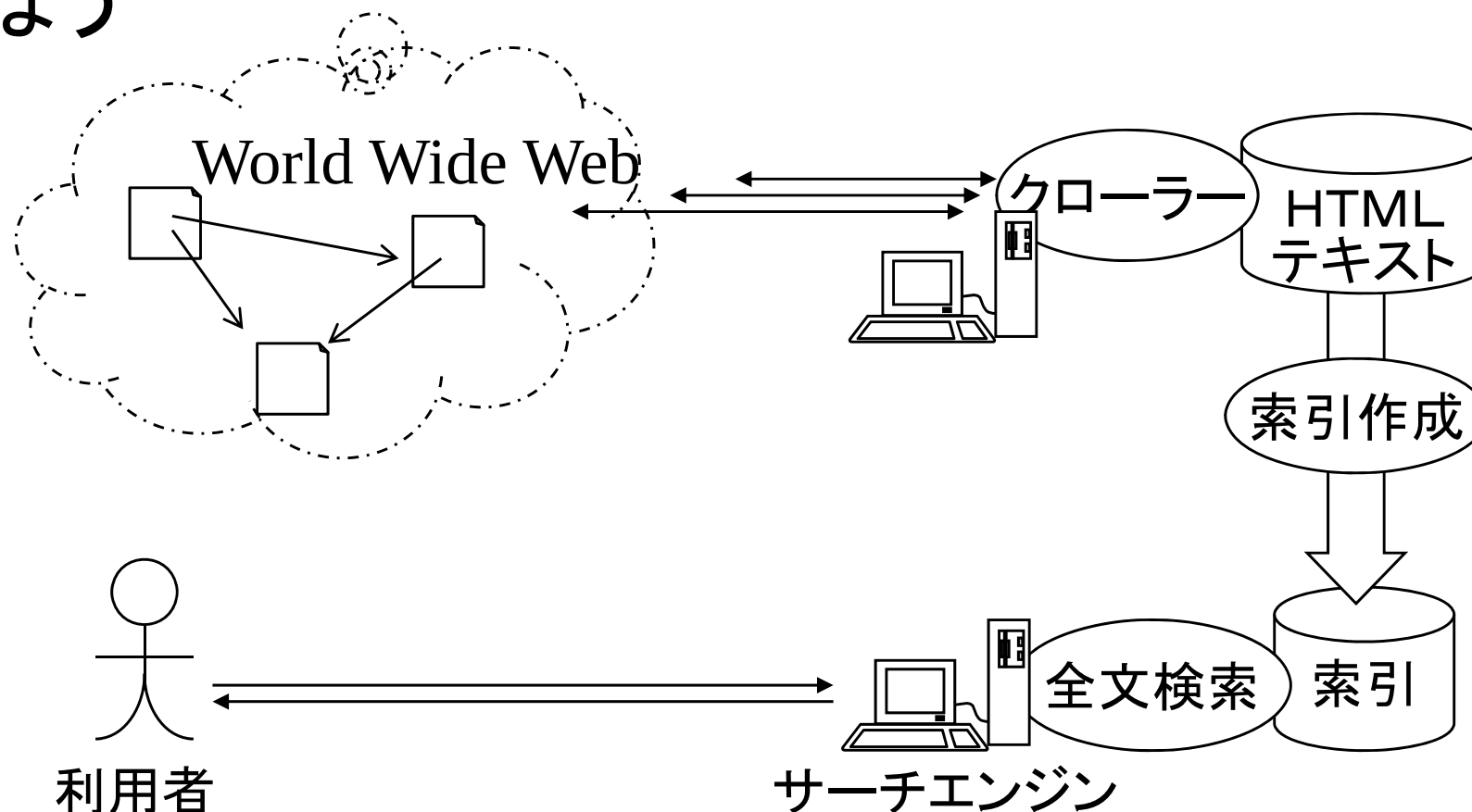
iモード

EZweb

Yahoo!ケータイ

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)

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

特徴ベクトル

属性

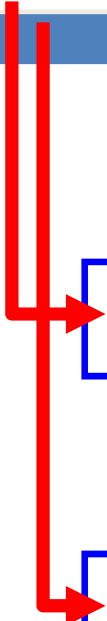
個体

	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

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

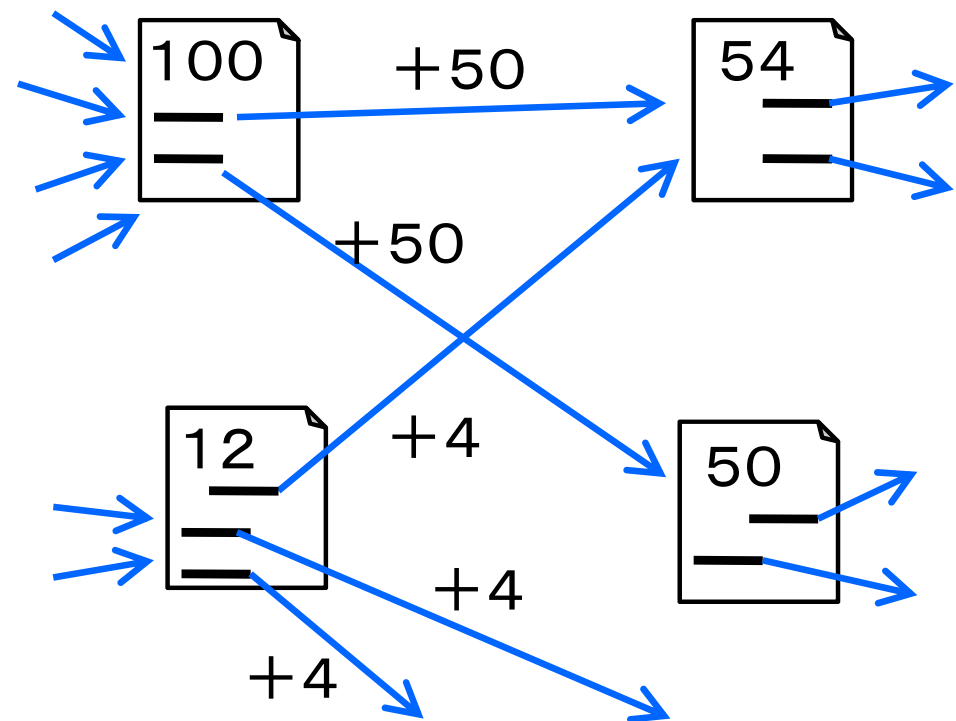
(1)

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」

$$R(u) = \sum_{v \rightarrow u} \frac{R(v)}{|B_v|}$$

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

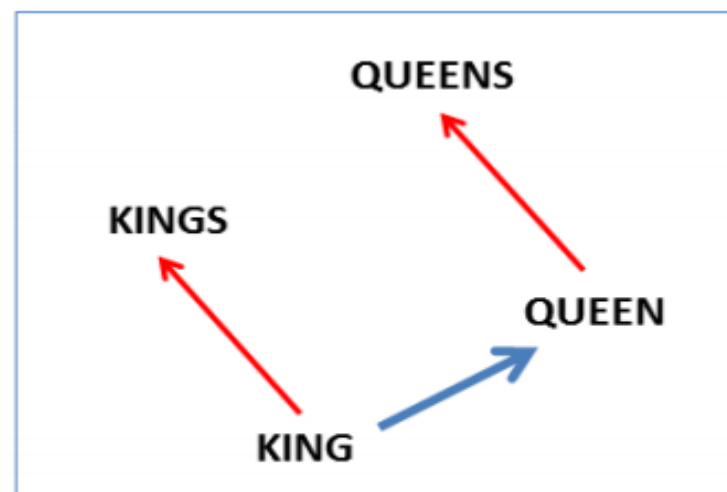
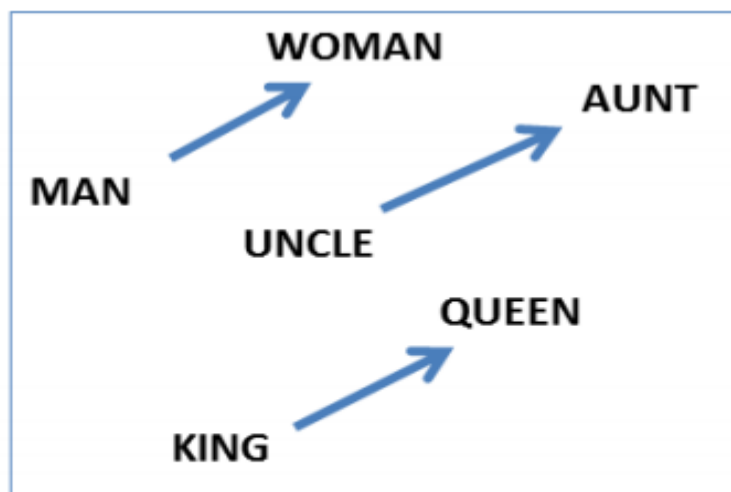


語 × 文書の行列

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

分布類似度

- 意味が似た語は似たコンテキストで出現
[Firth57]
- 語 × 文書行列で各語の行ベクトルを語の意味表現と考える→語の類似度計算が可能
- この考え方をNNで実現したものがWord2Vec



(Mikolov et al., NAACL HLT, 2013)

情報推薦

$$(4 \times 0.9 + 5 \times 0.7) / 2 = 3.6$$

	商品1	商品2	商品3	商品4	
ユーザ A	5	1	?	2	ユーザ A との類似度
ユーザ B		4	1	3	—0.9
ユーザ C	2	5	1		—0.3
ユーザ D	5	2	4	2	0.9
ユーザ E	5		5	3	0.7

フォークソノミー (folksonomy)

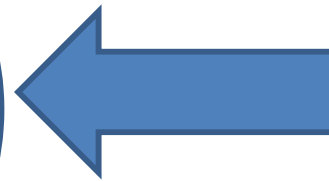
- folks (民衆) + taxonomy (分類法)
- ユーザによるウェブ上の情報へのタグ付け (分類)
- 共同作業による分類. タグの検索やタグを付けた人, その人がつけた他のタグを調べることができる.

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)

情報信頼性

批判的思考 (critical thinking)

合理的で内省的な思考
(他者ではなく自分の思考
過程を批判的に吟味する)



?

10秒息をとめることが
できれば、コロナウイ
ルスに感染していない

関連情報

情報分析システムWISDOM

NICT情報信頼性プロジェクト(2006~2010;プロジェクトリーダー:黒橋)



分析結果の
注目ポイント

主要・対立・
対比文

トピックの定義

発信者ごとの
意見の分布

肯定意見

否定意見

情報分析 システム WISDOM (NICT)

関連 キーワード

関連キーワード

重要キーワード

歯磨き:192 歯垢:119 歯茎:70 汚れ:85 ツルツル:58

～歯ブラシ

電動歯ブラシ:783 音波式電動歯ブラシ:8 音波歯ブラシ:51

歯～

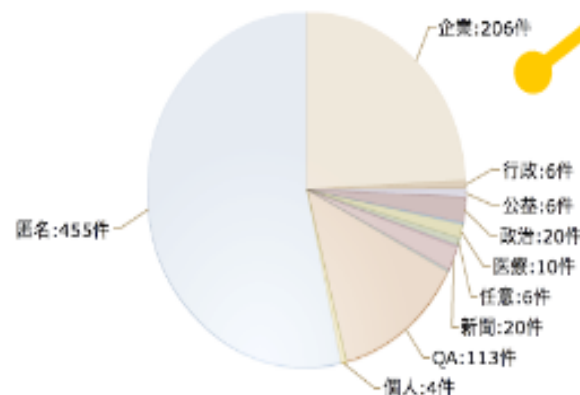
歯の表面:7 歯の隙間:4 歯医者:130 歯間ポケット:10 歯肉:10

その他

音波振動:33 臭:120 手磨き:22 毛先:32 白歯:33 効果:72 ブラウン:39

替えブラシ:20 ブラッシング:25 歯磨き粉:20 ドルツ:14 電動ハブラシ:12

発信者の分布



発信者の分布

主な発信者・主な意見

発信者クラス	主な発信者	肯定的な意見	否定的な意見
	DH:3件	...電動歯ブラシとは違い、1分間に40,000回以上の超高速で毛先が振動するハブラシ。...	...シ、電動歯ブラシを試用テストしましたが、どれも満足のものではありませんでした。...
企業	OGAWA Kenichi:6件 株式会社エヌディエス・NDS Corp.:5件	気軽に買える電動歯ブラシとしては性能、品質が良くお薦め品です。...	...歯ブラシ本体や、ブラシの先の形状・大きさによりうまくみがけない部分があるためです。...
政府・行政	No. New Nanga:6件 和歌山県火災予防保安協会:1件	...の歯磨き粉の方が すっきり感がある、口臭予防効果も高いので、やっぱりそっちを... ...ブラシも安く、しかも軽く当てただけなのでブラシの痛みも少ないのではと思っています。...	電動歯ブラシにとって替えブラシがないのは意味がないと思うのですが。...
公益法人等	国立情報学研究所:1件 株式会社京都青倉:1件	...める為の製品は、歯磨剤をはじめ洗口剤、電動歯ブラシや音波歯ブラシなど実に様々です。...	...シだけでお口の中を隅々まで磨くのは難しいですし、長期に使用することにより歯が余分に...
政治団体	日本諸おもしろ発見隊:12件 会津若松市立第四中学校:1件 医療法人社団スマイル会:1件	...音波ブラシやインタブレードなどは矯正のプラケットが入っていても使用可能でしょう。...	...している状態で、電動歯ブラシを使うのは装置が壊れる危険があるし、いくら強力な歯ブ...
大学・学会	兵庫歯科大学歯学部:1件 大学病院医療情報センター:1件	...電動歯ブラシにはまり、いろいろなメーカーを試しましたが、やはりブラウンが良いです。...	...初めての電動歯ブラシですが、ヘッドがゴツク奥歯の外側が磨きにくいような気がします。...
医療機関	Tsukaguchi Dental Clin.:5件 中川駅前歯科クリニック:2件 小峰歯科医院:1件	電動歯ブラシは、手にはない独特の動きでみがきにくい部分をみがくことができます。...	...千円歯ブラシよりも歯垢(プラーク)の除去効果はやや高いようですが(下関)、電動歯ブ...
任意団体	Fate hollow:4件 EOS Kiss Digital:2件	静かで小さな電動歯ブラシが人気！ もう電動歯ブラシは手放せない！	...単そうに見えて電動歯ブラシの使用方は、手用歯ブラシよりも難しいといわれています。...

企業 品質が良く お薦め品です

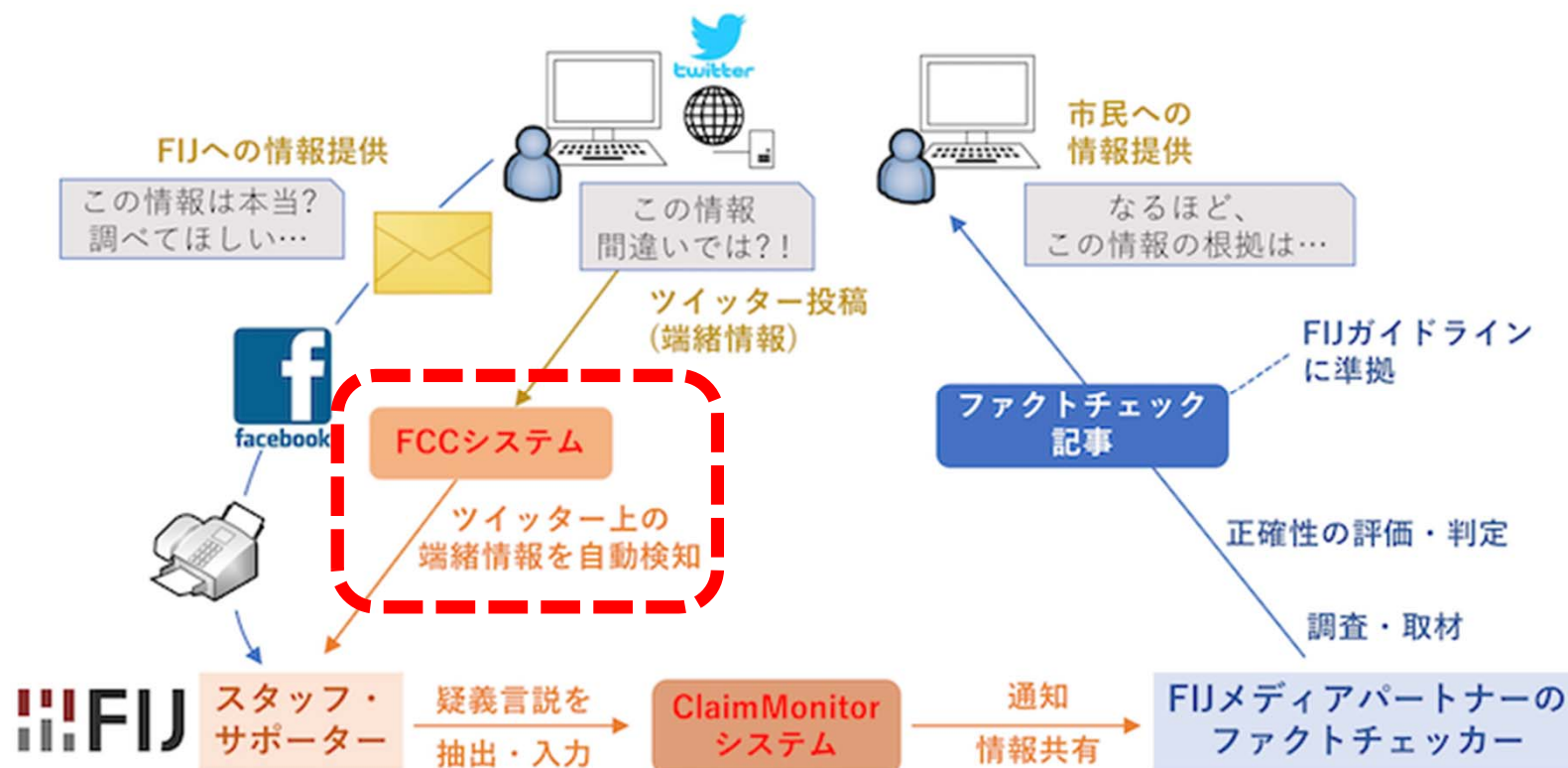
主な発信者と 主な意見

医療機関

簡単そうに見えて電動歯ブラシの使用方は、手用歯ブラシよりも難しいといわれています。

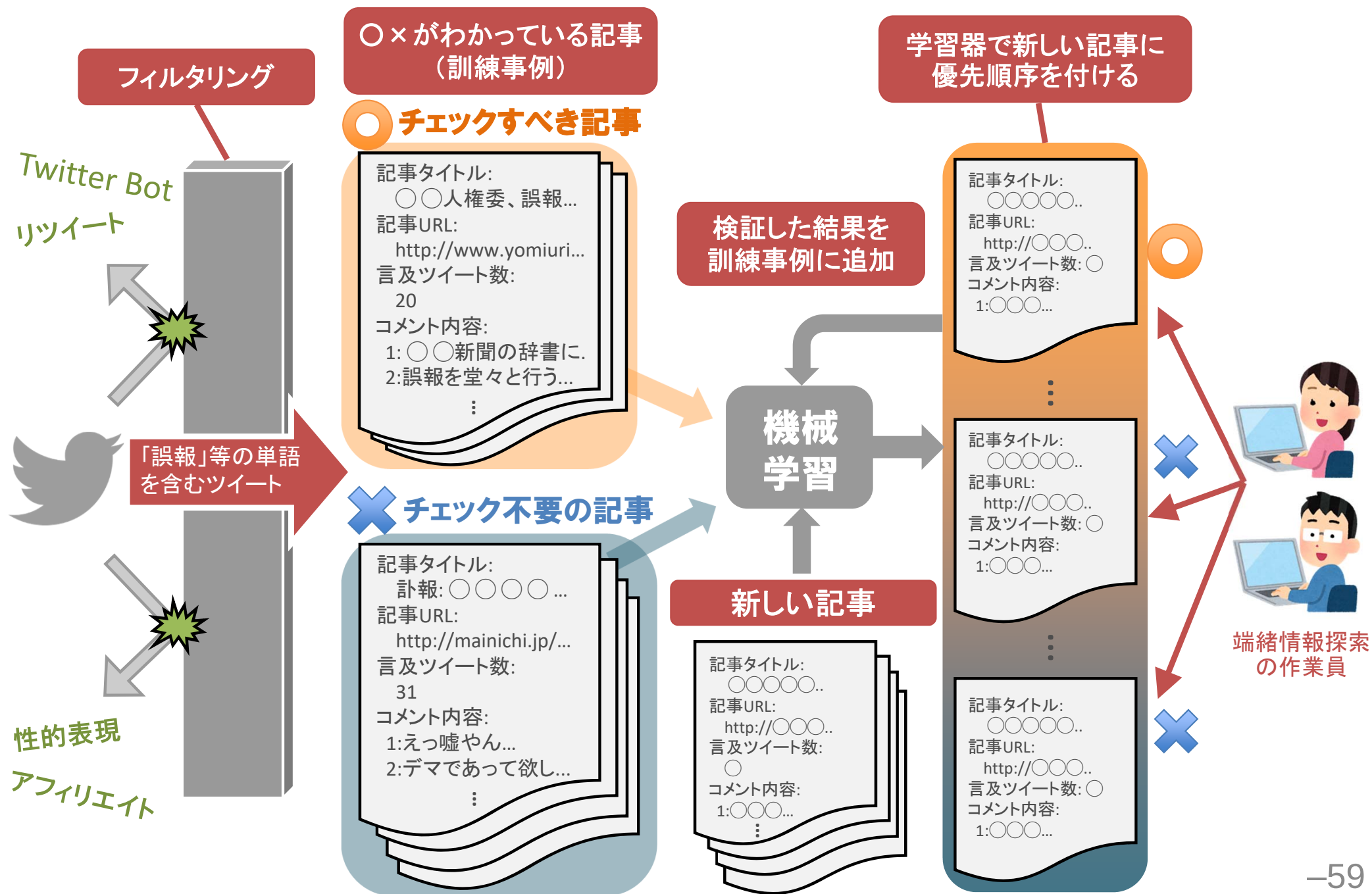
電動歯ブラシは、手にはない独特の動きでみがきにくい部分をみがくことができます。

ファクトチェック: 社会に広がっている情報・ニュースや言説が事実に基づいているかどうかを調べ、そのプロセスを記事化して、正確な情報を人々と共有



自然言語処理・機械学習を使った記事の絞り込み

東北大学 乾研究室



“

神経科学者のRodolfo Llinasによると、もし少なくとも10秒、咳や不快感なしに息をとめることができるならば、肺の繊維症が発生しておらず、コロナウイルスに感染していない。
—WhatsAppの音声投稿

”

Newtral.es (スペイン) による検証結果 **誤り**

Rodolfo Llinasは、そのような発言をしていない、効果的である証拠がない、と主張した。

ファクトチェック記事 (Newtral.es 2020.3.11)

“

ノーベル賞受賞の本庶教授が「武漢の研究所で4年間働いた」「(このウイルスは) 中国が造ったもの」などと発言
—4月中旬、フェイスブック(英文)などで拡散

”

インファクトによる検証結果 **虚偽**

本庶特別教授は新型コロナウイルスについて中国が造ったものだという発言をした事実はない。所属する京都大学、本庶教授の双方が否定した。本庶教授が中国・武漢の研究所に在籍した事実もない。

ファクトチェック記事 「ノーベル賞受賞者の本庶教授が中国の人造ウイルス説に言及」との虚偽情報がインドで拡散 **虚偽** (INFAC 2020.4.28)

ま と め

- LATCH: 5つの整理棚
 - Location, Alphabet, Time, Category, Hierarchy
- 人為分類 vs. 類型分類(類似性)
 - 動植物分類
 - 図書の分類法
 - ことばの分類
 - 情報検索、情報推薦
- 情報信頼性、批判的思考

6/16, 6/23: 演習

● 6/16(火)

- 演習課題: 図書をNDC(日本十進分類法)に基づいて分類する
KULINEの分類検索機能を用いて資料を探す
図書分類法とフォークソノミーを比較する
- 事前課題: **あり**

● 6/23(火)

- 演習課題: NDLC(国立国会図書館分類表)を用いて
資料を探す
NDCとNDLCの分類体系を比較する
資料探索に分類を用いる意義を考察する
- 事前課題: **あり**

【課題B-1】に解答する

【課題B-1】の解答内容は、来週のグループワークで使います。

- 解答・提出：PandA>テスト・クイズ>
20200616_第6回「情報の組織化と検索：理論と実
際」事前課題（6/15〆切）
- 締切：**2020/6/15（月） 12:00**

提出方法

- PandA「大学図書館の活用と情報探索」のページ

PandA

マイワークスペース ▾ [2018前期火4]大学図書館の活用と情報探索 ▾ 大学図書館の活用と情報探索 <2015後期火4> ▾

ホーム 🏠

スケジュール 📅

お知らせ 📢

授業資料 📖

課題 📋

Zoom (KU License) 📺

① テスト・クイズ 📝

フォーラム 🗨️

メディアギャラリー 🖼️

BookRoll 📖

ヘルプ ?

🔄 [2020前期火4]大学図書館の活用と情報探索: テスト・クイズ

テスト

テストを受験

受験可能なテストの一覧は

タイトル 📄

① 20200616_第6回「情報の組織化と検索：理論と実際」事前課題（6/15〆切）」をクリック

[20200512_第1回授業アンケート](#)

[20200519_第2回授業アンケート](#)

N/A

2020-5-26 11:59 午前

提出済みテスト

すべての提出物・得点を表示 |

タイトル	統計情報	記録済み点数	フィードバック利用可能日時	個々の得点
------	------	--------	---------------	-------

提出方法

今日の授業はどのくらい役に立ちそうですか？

[選択をリセット](#)

② 解答を記入

問題 3 / 4

今日の授業の進み具合について5段階で選択してください。

	1.とても速かった	2.速かった	3.適切だった	4.遅かった	5.とても遅かった
今日の授業の進み具合をどう感じましたか？	☐	☐	☐	☐	☐

[選択をリセット](#)

問題 4 / 4

今日の授業について一番重要だと思ったことなど、感想を自由にお書きください。

※注意※ここで質問しても直接回答ができません。回答が必要な質問は、フォーラムに記入するか、メールでご連絡ください。

(最大文字数: 60,000 文字)

文字数をカウント

[リッチテキストエディタを表示・非表示](#)

③

保存

終了

採点のために提出

[保存]...提出はせずに、回答の保存のみ行います。

[終了]...回答を保存せずにテストを終了します。

[採点のために提出]...回答を提出し、テストの採点を行います。

③ 「採点のために提出」をクリック
※押し忘れに注意