情報検索

「さがす」の情報科学 [吉川2007]

概要

世の中は情報を再利用（探し出して改善）することで進歩してきました．今日はこの「探し出す」事の歴史的推移や，最近では当たり前になっている検索エンジンの技術要素について眺めてみましょう．

キーワード: 情報検索，自然言語処理，類似度

當間愛晃　@　琉球大学工学部情報工学科
E-mail:　tnal@ie.u-ryukyu.ac.jp
Web:　http://www.eva.ie.u-ryukyu.ac.jp/~tnal/
ブログ:　http://ie.u-ryukyu.ac.jp/tnal/

イントロダクション

問: これまでに勉強や趣味のために模索/詮索/検索/物色/etc.をした事がありますか？

・調べようとしたモノは何ですか？
・どのように調べましたか？
・望んでいたモノは見つかりましたか？

本日のお題：情報検索

授業の進め方

＞授業中

・グループ作業

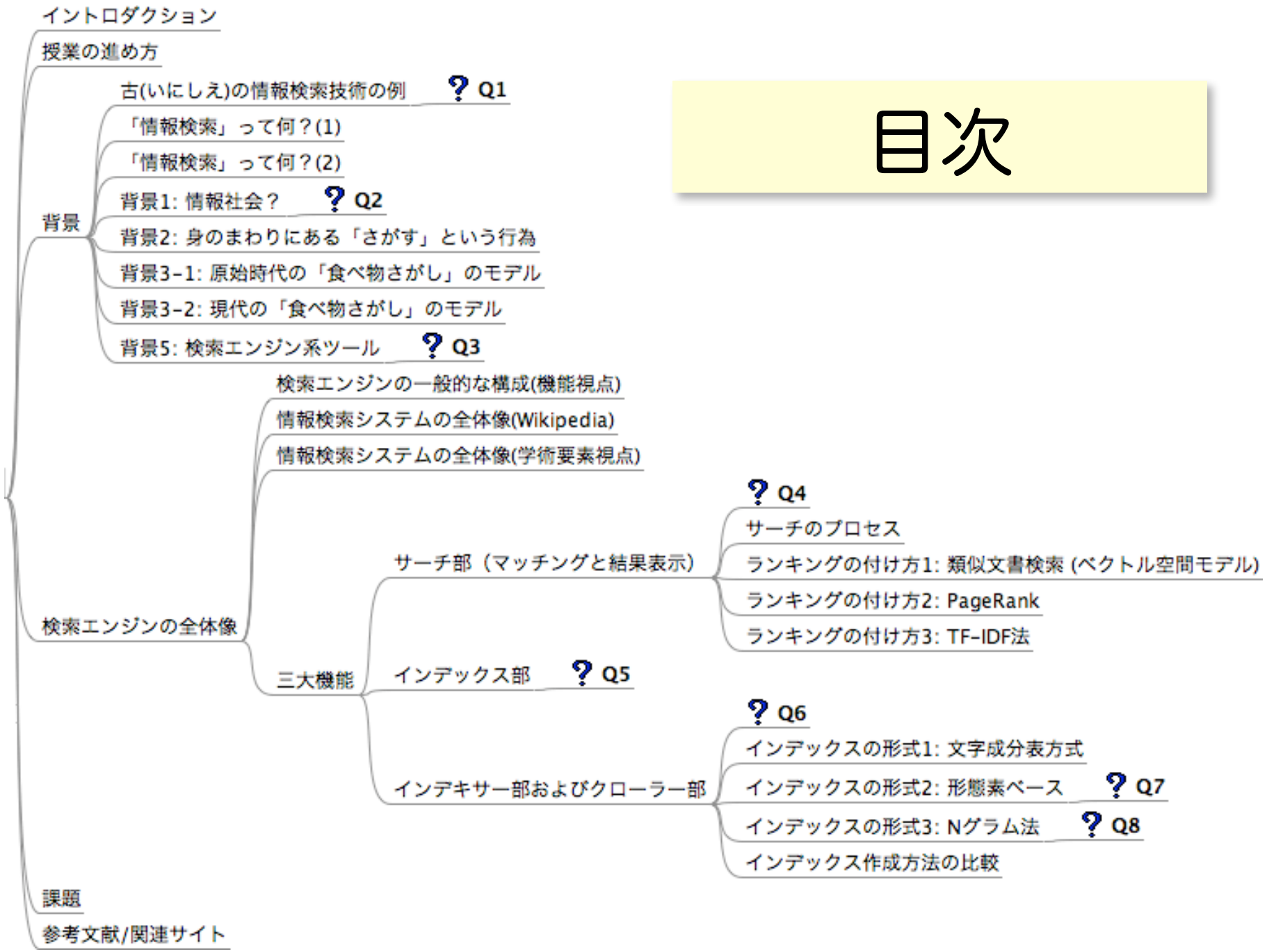
各質問毎に，「解説→グループ討論→全体討論（報告会）」を行う．

＞提出課題

・個人作業

課題についてレポートを作成し，当日回収を予定．

目次


イントロダクション
授業の進め方
背景
古(いにしえ)の情報検索技術の例
Q1
「情報検索」って何？(1)
「情報検索」って何？(2)
背景1: 情報社会？
Q2
背景2: 身のまわりにある「さがす」という行為
背景3-1: 原始時代の「食べ物さがし」のモデル
背景3-2: 現代の「食べ物さがし」のモデル
背景5: 検索エンジン系ツール
Q3
検索エンジンの全体像
検索エンジンの一般的な構成(機能視点)
情報検索システムの全体像(Wikipedia)
情報検索システムの全体像(学術要素視点)
三大機能
サーチ部（マッチングと結果表示）
Q4
サーチのプロセス
ランキングの付け方1: 類似文書検索 (ベクトル空間モデル)
ランキングの付け方2: PageRank
ランキングの付け方3: TF-IDF法
インデックス部
Q5
インデキサー部およびクローラー部
Q6
インデックスの形式1: 文字成分表方式
インデックスの形式2: 形態素ベース
Q7
インデックスの形式3: Nグラム法
Q8
インデックス作成方法の比較
課題
参考文献/関連サイト

古(いにしえ)の情報検索技術の例

＞ユニターム・カード

ユニタームカード（Uniterm card）は主題語毎に作成したカードに文献の番号を書き込んだものである。たとえば「インドにおける医薬産業」についての文献を探したいときは「インド」と「医薬産業」のカードを選び、2つのカードに共通に書かれている文献番号を探せばよい。現在のオンライン検索システムの基礎原理はここから発している。 *（出典: フリー百科事典『ウィキペディア（Wikipedia）』）*

Q1: これで何か問題があるだろうか？
あるとしたら何が問題なのか？

「情報検索」って何？(1)

＞情報検索

情報検索（じょうほうけんさく）とは、コンピュータを用いて**大量のデータ群**から**目的に合致したものを取り出す**こと。検索の対象となるデータには**文書や画像、音声、映像、その他さまざまなメディア**やその組み合わせとして記録されたデータなどが含まれる。インターネットの発達により検索はインターネットを介して行われることも多いが、ここでは情報を検索するためのコンピュータ側における仕組みを記述している。*（出典: フリー百科事典『ウィキペディア（Wikipedia）』）*

「情報検索」って何？(2)

[論点1] 大量のデータ群
- どのぐらい？
- 大量だと何が問題になるの？

[論点2] 目的に合致したものを取り出す
- 「目的に合致したもの」？とはどのように定義できるか？
- どうやって見つけるの？

[論点3] 文書や画像、音声、映像、その他さまざまなメディア
- それぞれのメディアでどうやって「目的に合致したもの」を見つけるの？（今回は「文書」）

情報を検索するためのコンピュータ側における仕組み
- それは利用者にとって十分に使いやすい？（討論予定）
- もっと便利な他の方法はないの？（討論予定）

背景1: 情報社会？ *(吉川2007,p.2)*

＞デジタルデータ量

米IDEによる2007年3月の調査結果

2003年: 5EB
2006年: 161EB
2007年: 255EB(見込み)

＞EB?

1EB (ExaByte) = 1,024PB
1PB (PetaByte) = 1,024TB
1TB (TeraByte) = 1,024GB

(参考) iPod nano 16GB
・写真14,000枚
・音楽4,000曲
・ビデオ16時間

＞検索システムの需要増加

『情報が少ない時代には、情報は一覧性・アクセス性が高いように十分整理・分類されて保管されていた。しかし、デジタル情報の爆発的増加に伴い、この一覧性とアクセス性が損なわれてしまったのである。』

Q2: 一覧性とアクセス性が損なわれると何が困るのだろうか？

背景2: 身のまわりにある「さがす」という行為 *(吉川2007,p.10)*

(例1)

○ ○さん，ほら，先日のA社との契約の時に作ったあの協議用の書類，あれに引用してあったデータの元のファイル，さがして持って来てくれるかな？

＞行為

・書類さがし.
整理保管されているか？
・引用元さがし.
引用元が書かれているか？
・引用元のファイルさがし.
整理保管されているか？

(例2)

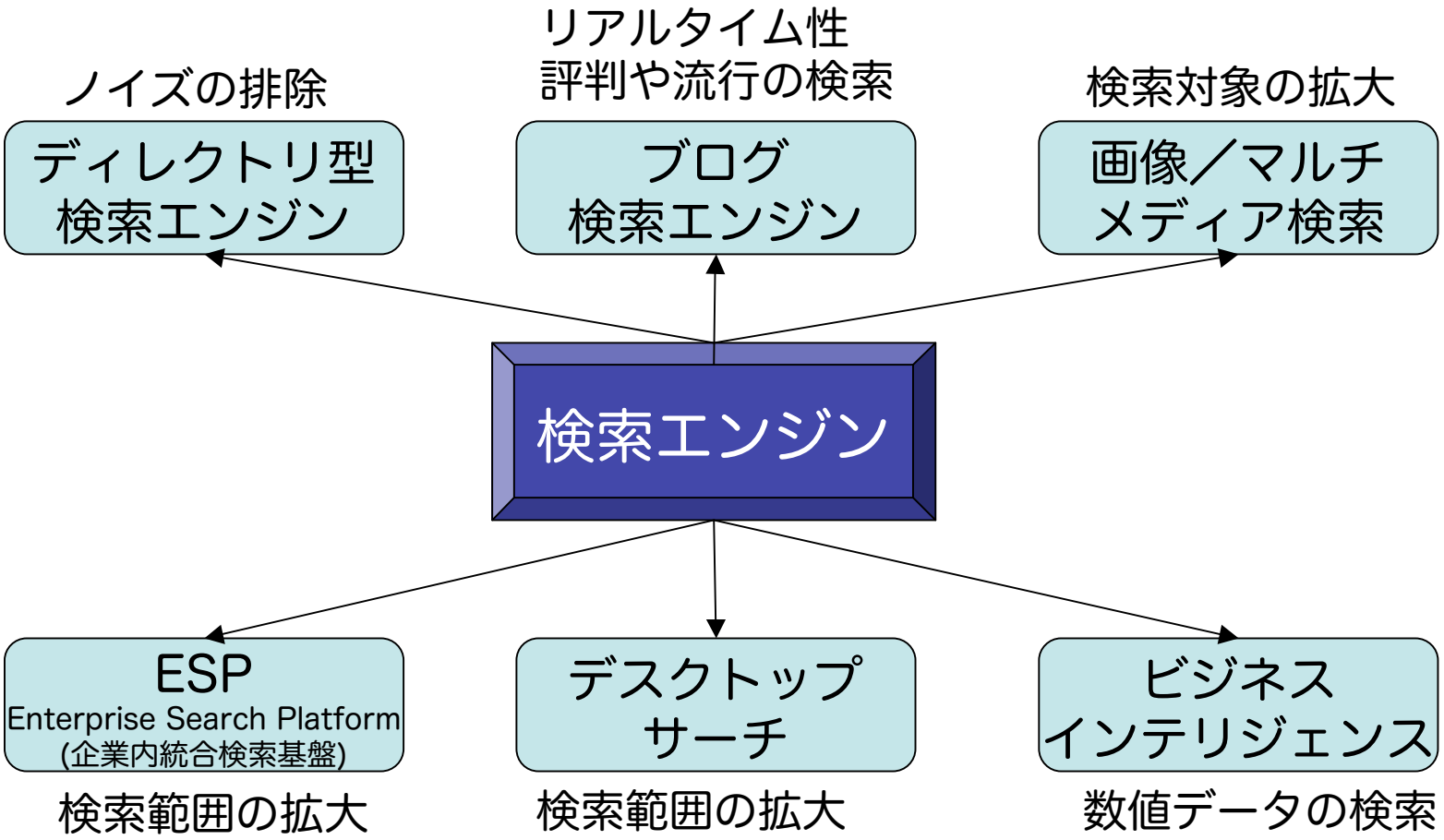
●●くん，君にはこのたび新規事業開発担当として新しい製品を開発してもらうことになった. ついては，君の頭を振り絞って今までに無い斬新なアイデアを出して欲しい. アイデアは企画書として簡単にまとめて来週までに提出する事.

＞行為

・企画書のイメージさがし.
どう書けば良いのか？
・アイデアさがし.
・競合他社との違いさがし.
関連事業にはどのようなものが？

指示そのものは「提案や企画書作成」だが，それを達成するには多種多様な「さがす」行為が必要.

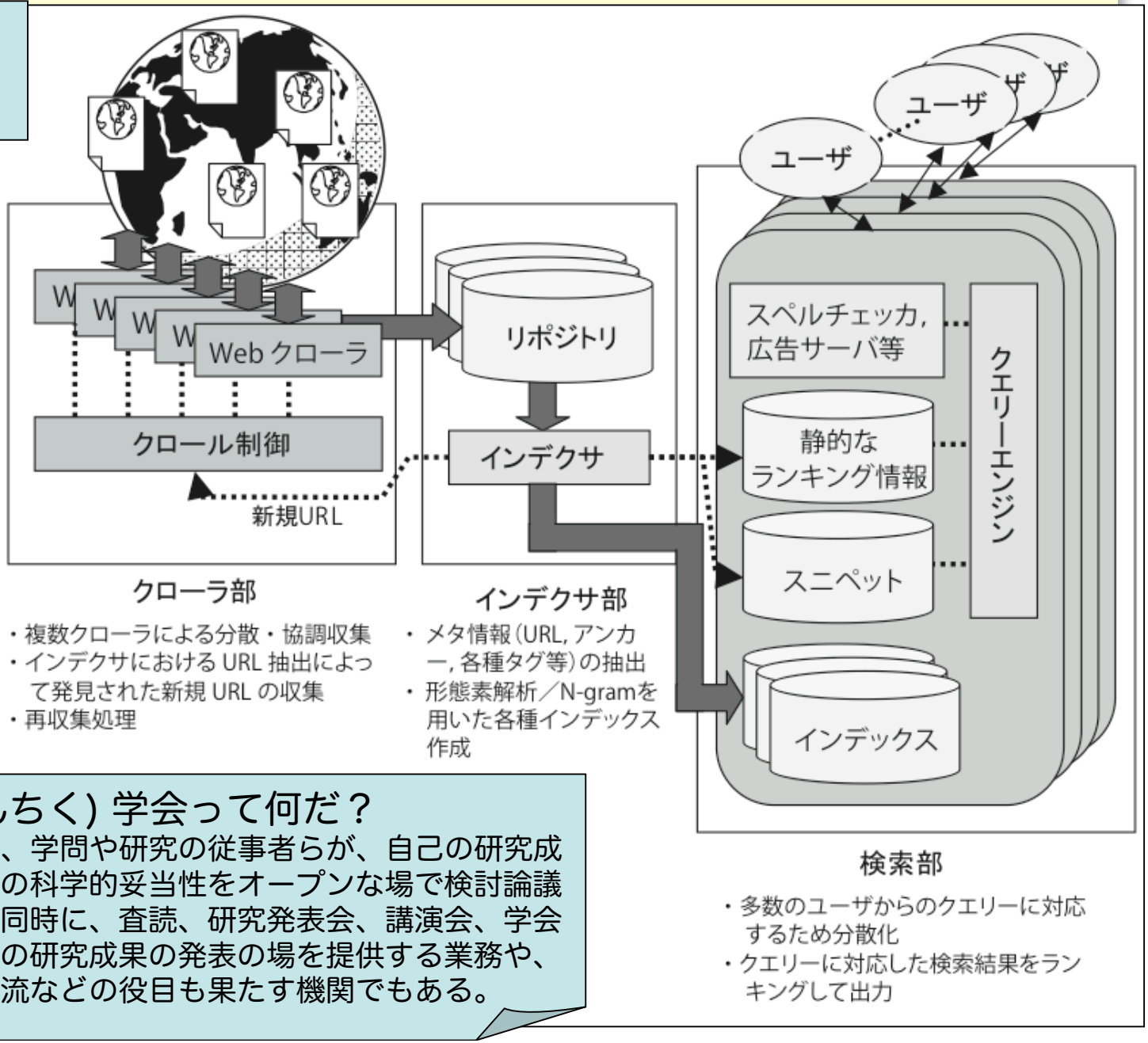
背景5: 検索エンジン系ツール *(吉川2007,p.70)*



Q3: それぞれ具体的にどのような例があるだろうか？

検索エンジンの一般的な構成(機能視点)

[山名 2005] 情報処理学会会誌「情報処理」より掲載



(うんちく) 学会って何だ？

学会（がっかい）は、学問や研究の従事者らが、自己の研究成果を公開発表し、その科学的妥当性をオープンな場で検討論議する場である。また同時に、査読、研究発表会、講演会、学会誌、学術論文誌などの研究成果の発表の場を提供する業務や、研究者同士の交流などの役目も果たす機関でもある。

情報検索システムの全体像(学術要素視点)

- データの管理および入出力のためのデータベース
- 文書データ処理のための自然言語処理や計算言語学
- 画像や音声を扱うための(デジタル)信号処理
- 認知心理学を背景とするパターン認識技術
- メタデータに関する考察の基盤となった図書館情報学
- 検索アルゴリズム設計や情報検索システムの評価尺度考案のヒントとして寄与した諸数学理論

[Wikipedia:情報検索]

- 大容量を高速度で処理するための並列分散処理技術

インデックスの形式1: 文字成分表方式
(吉川2007,p.100)

文書中に現れるすべての文字の種類を記録しておく方式.

	あ	い	う	え	お	…	検	…	索	…
文書1	0	0	0	1	1		1		1	
文書2	1	1	0	1	0		0		1	
文書3	1	0	1	1	1		0		0	
文書4	0	0	0	0	1		1		1	
…										

(うんちく) 日本人涙目！
単語がスペースで区切られている英語圏と比較して，単語をどうやって抽出するかというレベルから研究せざるを得なかった.

インデックスの形式2: 形態素ベース
(吉川2007,p.103)

形態素＝文章を文字列に区切っていったときの意味を持つ最小の単位.
形態素解析＝文章を形態素毎に切り出すこと．品詞特定.

追加工夫例
・助詞（てにおは等）はあまり意味を持たないので削除.
・活用形（関する，関して，関した）を同義語とみなす.

「検索に関する技術を解説した本を書く」

→「検索」「に」✗「関する」「技術」「を」✗「解説」「した」✗「本」「を」✗「書く」

Q7: 形態素解析後の「追加工夫例」は計算効率を高める工夫の例である．他にどのような工夫が考えられるか？

インデックスの形式3: Nグラム法
(吉川2007,p.107)

文中から1文字ずつずらしたN文字の並びを順に取り出してインデックスを作る，N文字インデックスと呼ばれる方式.

「文書検索のノウハウを説明」

文字成分表方式の2文字単位の場合

→「文書」「検索」「のノ」「ウハ」「ウを」「説明」

2グラム方式（N=2）の場合

→「文書」「書検」「検索」「索の」「のノ」「ノウ」「ウハ」「ハウ」「ウを」「を説」「説明」

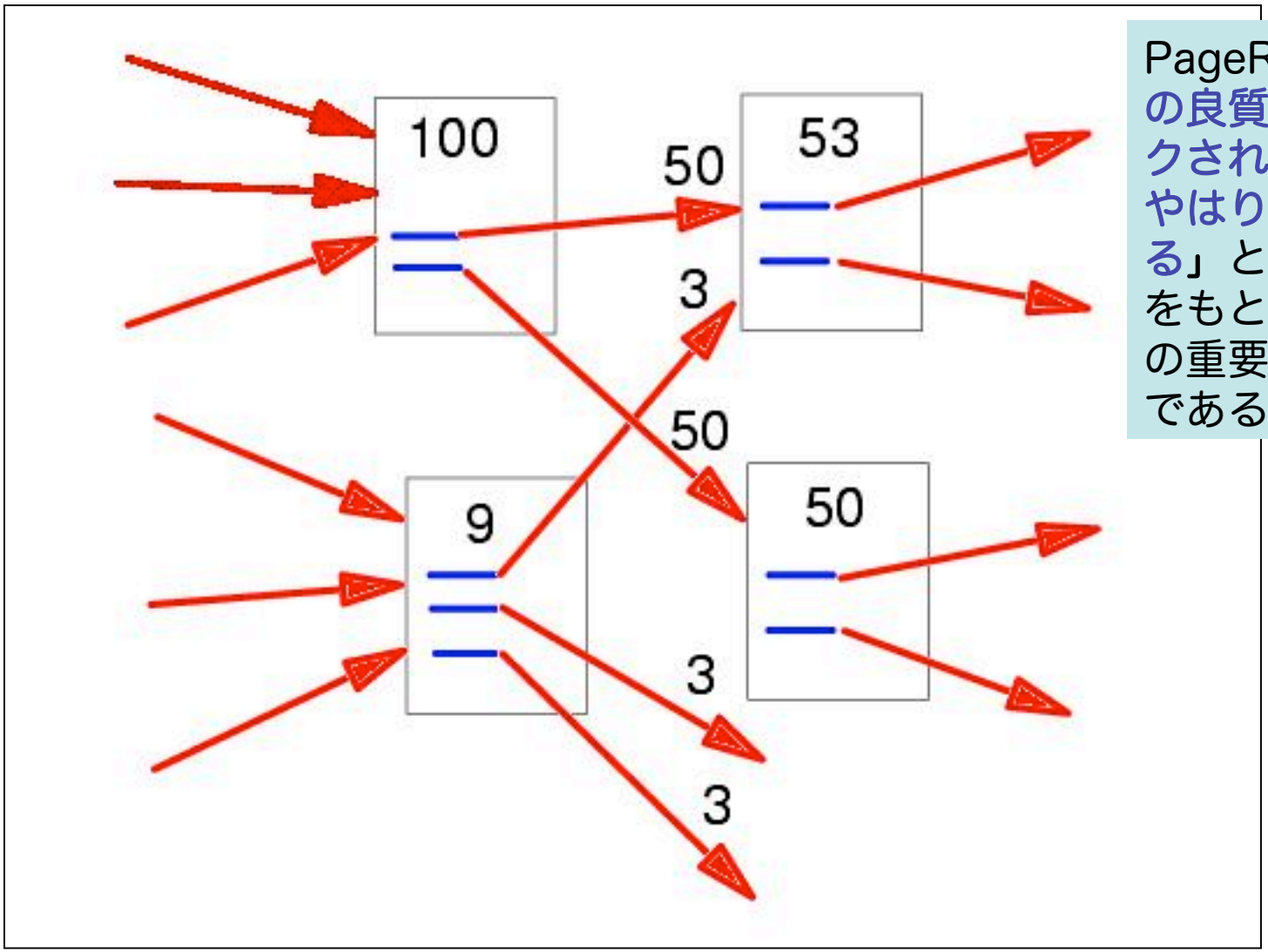
Q8: この違いが検索結果にどのような影響を及ぼすだろうか？

インデックス作成方法の比較
(吉川2007,p.114)

	形態素	Nグラム
ノイズ	少ない	多い
検索漏れ	あり	少ない
インデックスの作成時間	遅い	早い
インデックスのサイズ	小さい	大きい
検索時間	早い	遅い
辞書	必要	不要
言語依存	する	しない

ランキングの付け方2: PageRank

Google の秘密 - PageRank 徹底解説のページより
http://homepage2.nifty.com/baba_hajime/wais/pagerank.html



PageRank は、「多くの良質なページからリンクされているページは、やはり良質なページである」という再帰的な関係をもとに、全てのページの重要度を判定したものである。

ランキングの付け方3: TF-IDF法

キーワードの出現度を示すTFと重要度を示すIDFのページより
http://www.searchengineoptimization.jp/seo_foundation/scorering/tfidf.html

TF-IDF法は、文字どおりTF(term frequency)という指標とIDF(inverse document frequency)という指標の2つの指標を用いたアルゴリズムです。TFとIDFの双方の指標を用いることで、キーワードに対する個々のWebページのスコアリングを、キーワードの重要性をふまえた上で行うことができます。

＞TF(term frequency)

TFとは、Webページ内でキーワードがどれだけ多く使用されているのかを示す指標です。キーワードを多く含むWebページほど、そのキーワードについて詳しく説明しているものと考えるものです。

＞IDF(inverse document frequency)

IDFとは、そのキーワードがどれだけの数のWebページで使用されているかを示す指標です。多くのWebページで使用されているキーワードより、少ないWebページで使用されているキーワードの方が重要性の高いものと考えるものです。

課題1

＞課題

これまでに「さがしもの」をして困った事例を一つ以上挙げて，以下の項目について説明せよ．

- 何をさがそうとしたのか．
- どのようにさがそうとしたのか．
- 何故困ったのか．
- (オプション) それを解決する方法を提案せよ．

＞提出方法

レポート用紙にまとめ，講義終了時に回収．講義中に聞けなかった／分からなかった点について質問等があれば，メール等で連絡ください！

課題2

＞課題

以下の論点（資料6,7ページ）について意見を述べよ.

[論点1] 大量のデータ群

- どのぐらい？
- 大量だと何が問題になるの？

[論点2] 目的に合致したものを取り出す

- 「目的に合致したもの」とはどのように定義できるか？
- どうやって見つけるの？

[その他] 感想

＞提出方法

レポート用紙にまとめ，講義終了時に回収．講義中に聞けなかった／分からなかった点について質問等があれば，メール等で連絡ください！

参考文献/関連サイト

＞Webページ

Wikipedia: 情報検索, 情報化社会, ユニターム・カード等
　　http://ja.wikipedia.org/wiki/情報検索

中川裕志: 講義「情報データベース論」（東京大学）
　　http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/infoDB/syllabus.html

人工知能のやさしい説明（人工知能学会）
　　http://www.ai-gakkai.or.jp/jsai/whatsai/

＞書籍（高校生でも読みやすいものを挙げています）

サーチアーキテクチャ - 「さがす」の情報科学
　　吉川日出行(編著), ソフトバンククリエイティブ出版; ISBN: 4797341033; (2007/9)

自然言語処理ことはじめ - 言葉を覚え会話のできるコンピュータ
　　荒木健治(著), 森北出版; ISBN: 4627828519; (2004/6)

マッチ箱の脳 (AI)―使える人工知能のお話
　　森川幸人 (著), 新紀元社 ; ISBN: 4883170802 ; (2000/12)

＞学術雑誌

[山名 2005] 山名・村田: 1. 検索エンジンの概要(<特集>検索エンジン2005-Webの道しるべ-), 情報処理学会学会誌「情報処理」, 46巻9号, pp.981-987, 2005年9月号.