大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

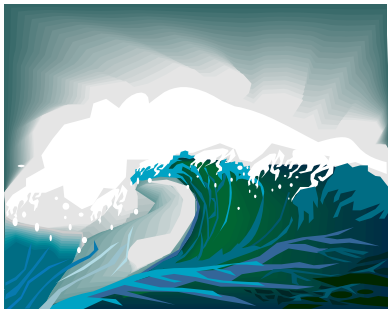
国立情報学研究所　2009年度市民講座　第8回　(2010.2.17)

膨大な文書の処理技術
～テキストの山を斬って見えてくるもの～

コンテンツ科学研究系
高須淳宏　**takasu@nii.ac.jp**

増大する電子テキスト情報　(1/2)

- Webのページ数
  - 日本のWebページ数(2004)　8,500万ページ[1)]
  - 世界の総Webページ数(2005)　115億ページ[2)]
- ブログ数
  - 国内のブログの総記事数(2008)　13億5,000万件[3)]

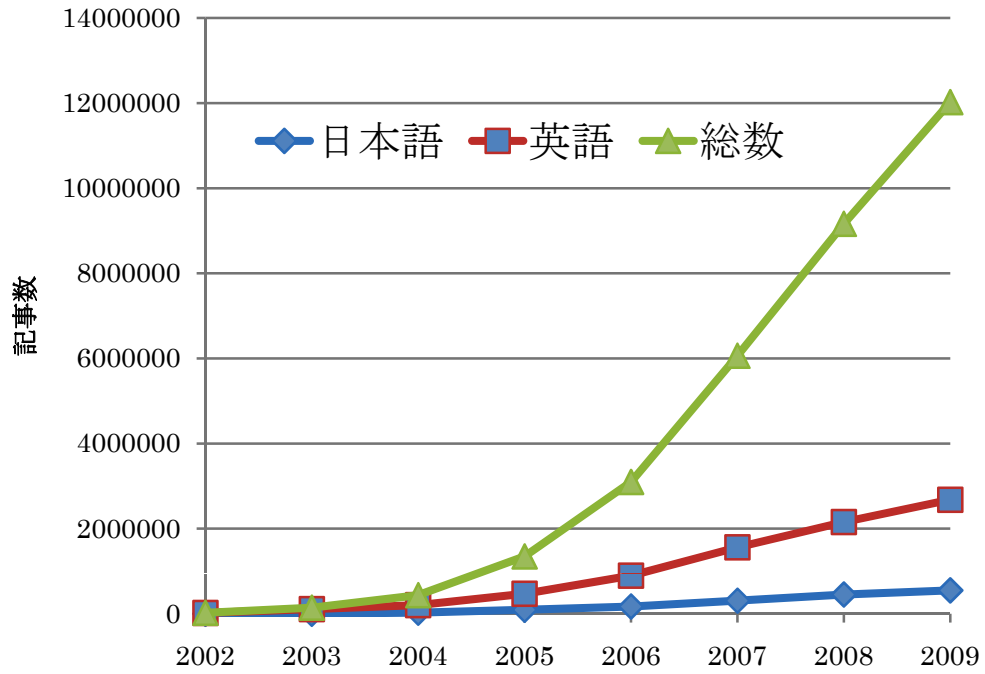


1)　総務省　情報通信政策研究所：
“インターネット概観統計集（平成18年改訂）”，　平成19年3月
2)　WWW05] A. Gulli, A. Signorini :
" The Indexable Web is More than 11.5 billion pages "
World Wide Web Conference, 2005.
3)　総務省　情報通信政策研究所：
“ブログの実態に関する調査研究の結果”，　平成20年7月.

増大する電子テキスト情報 (2/2)

- ウィキペディア　百科事典　(2009.12)
  - 日本語記事数　63万5千記事
  - 英語記事数　300万記事
  - 全記事数　1,400万記事

「Wikipedia 多言語の統計」より



情報のテキスト情報を効率よく活用する技術にむけて

- このような情報空間はあまり整理整頓のされていない図書館のようなもの
- 必要な情報が含まれている可能性が高いが、引き出すのも大変
  - 知的活動のうち30%を情報を探すことに費やすという調査報告も
- 情報を効率よく利用するテキスト処理技術にむけて
  - 探す： 情報検索、Web検索
  - 情報を見出す： 情報分析、テキストマイニング
  - できるだけ集める： 情報統合
- この講座では
  - 「コンピュータはどのようにテキストを扱か」という観点からテキストマイニング研究を紹介

概要

- コンピュータによるテキスト処理の概要
  - 標準的なテキストの表し方
- コピーコンテンツと文字列マッチング
  - 大学生のレポート作成事情
  - スパムブログをみつける
  - コンピュータにテキストはどこまで生成できるか？
  - 同一の事件を扱うニュース記事をまとめる
- 潜在トピックの抽出
  - 関連情報を一括して集める
  - ホットトピックをみつける

コンピュータはテキストをどうように扱うか (1/3)

例文:市民講座概要

現在、私たちは、インターネットを介して膨大な量のテキストにアクセスできるようになっています。また、電子メールやオンラインニュースなど、日々生成されるテキストの量も多く、これらのテキスト情報を上手に活用することが求められています。本講座では、まず、大量テキストデータを処理するための、最新の テキストマイニング技術を紹介します。インターネット上のテキスト情報の特徴のひとつに、大量の類似情報が含まれていることがあげられます。そしてこれら の最新技術の紹介に続いて、これらの類似テキスト情報を効率良く検出するための私たちの取り組みとその技術の応用事例を紹介します。

コンピュータはテキストをどうように扱うか (2/3)

現在、私たちは、インターネットを介して膨大な量のテキストにアクセスできるようになっています。

形態素解析

現在	名詞-副詞可能
私	名詞-代名詞-一般
たち	名詞-接尾-一般
は	助詞-係助詞
インターネット	名詞-一般
を	助詞-格助詞-一般
介す	動詞-自立
て	助詞-接続助詞
膨大	名詞-形容動詞語幹
だ	助動詞
量	名詞-一般
の	助詞-連体化

テキスト	名詞-一般
に	助詞-格助詞-一般
アクセス	名詞-サ変接続
できる	動詞-自立
よう	名詞-非自立-助動詞語幹
に	助詞-副詞化
なる	動詞-自立
て	助詞-接続助詞
いる	動詞-非自立
ます	助動詞

ChaSen: http://chasen-legacy.sourceforge.jp/

コンピュータはテキストをどうように扱うか (3/3)

現在、私たちは、インターネットを介して膨大な量のテキストにアクセスできるようになっています。また、電子メールやオンラインニュースなど、日々生成されるテキストの量も多く、これらのテキスト情報を上手に活用することが求められています。本講座では、まず、大量テキストデータを処理するための、最新の テキストマイニング技術を紹介します。インターネット上のテキスト情報の特徴のひとつに、大量の類似情報が含まれていることがあげられます。そしてこれら の最新技術の紹介に続いて、これらの類似テキスト情報を効率良く検出するための私たちの取り組みとその技術の応用事例を紹介します。

Bag-of-word

単語	出現回数
テキスト	7
情報	4
技術	3
紹介	3
量	2
類似	2
インターネット	2
最新	2
大量	2
アクセス	1

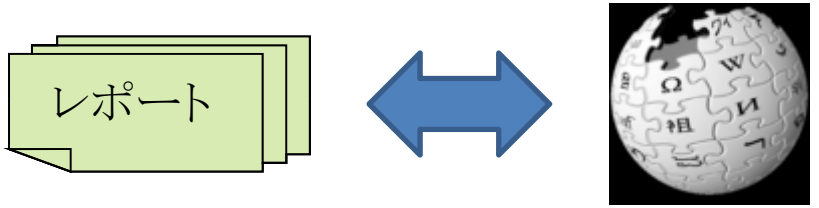
単語	出現回数
オンライン	1
データ	1
ニュース	1
マイニング	1
メール	1
応用	1
活用	1
検出	1
効率	1
講座	1

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

コピーコンテンツと文字列マッチング

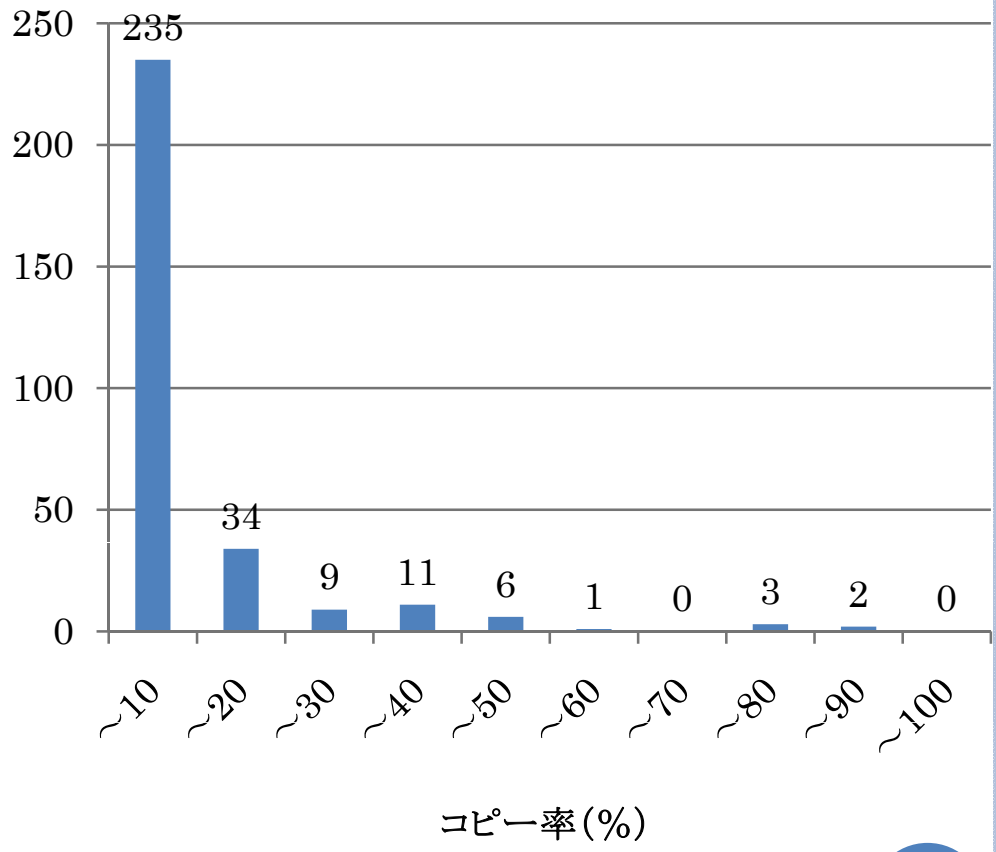
データからみる大学生のレポートの書き方

- 大学1、2年生を対象とした情報のクラスで提出された300名のレポート
- Wikipedia+αの日本語の全記事と共通する部分を検出
- レポートにおけるWikipediaに表れる語が使われている割合
  - 最小 61.4%　平均　95.2 %　最大　100%
- レポートにおいてWikipediaからコピーしたテキストの割合
  - 最小 0%　平均　7.2 %　最大　87.3%



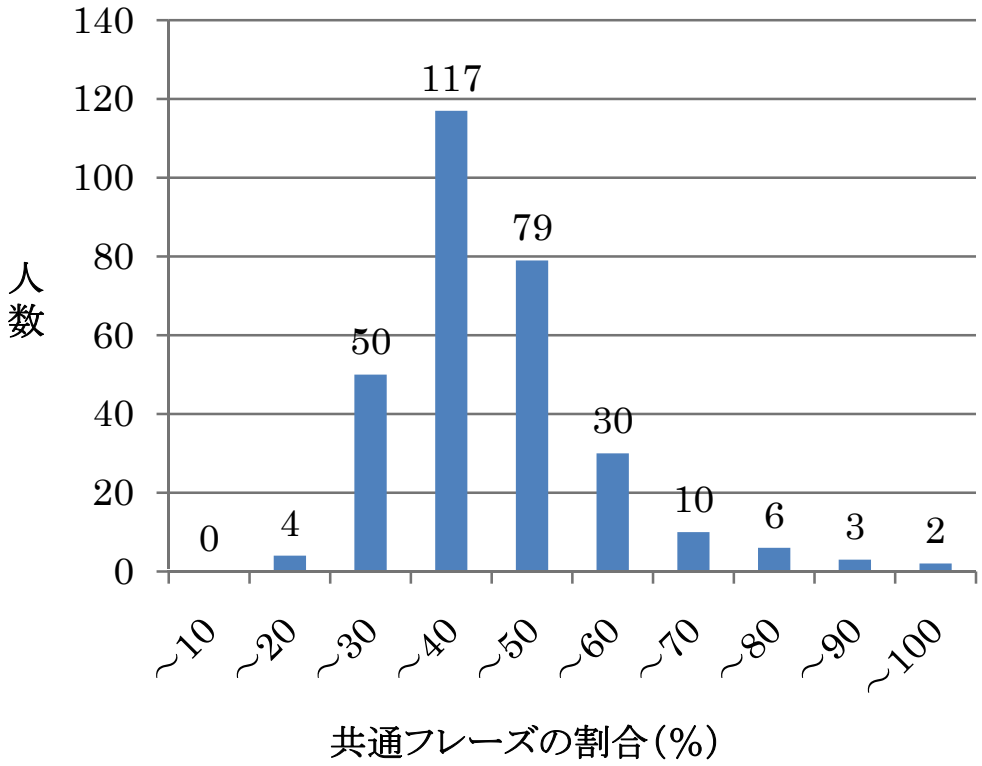
WIKIPEDIA＋からのコピーの割合

- Wikipedia＋αの記事と60バイト以上の長さの共通文字列のレポート全体に対する割合
- 多くの学生はまじめにレポートを作成しているようにみえる
- 数名の学生はかなり大胆にコピーを行っている



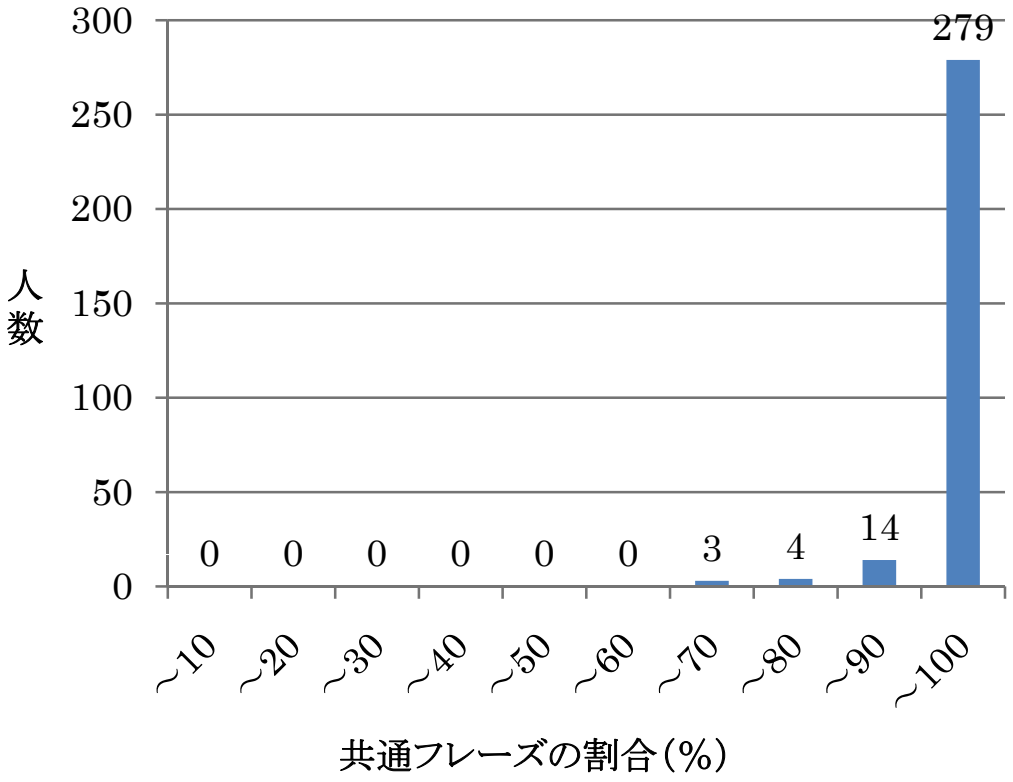
WIKIPEDIAの記事と共通するフレーズの割合

- Wikipedia＋αの記事と30バイト以上の長さの共通文字列のレポート全体に対する割合
- Wikipdeiaで使われているフレーズが比較的使われている



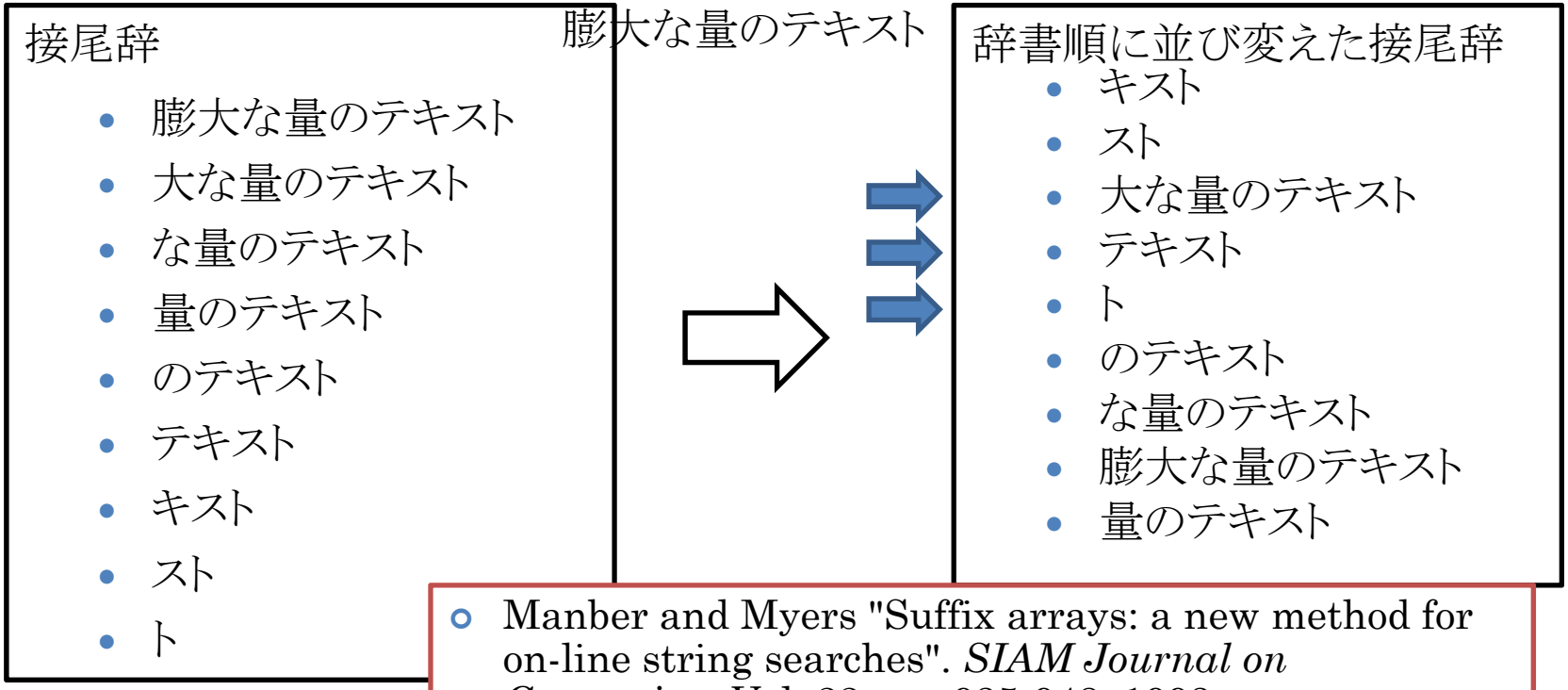
WIKIPEDIAの記事と共通する単語の割合

- Wikipediaの記事と10バイト以上の長さの共通文字列のレポート全体に対する割合
- Wikipediaに使われる単語と共通のものが多く使われている。
- 3つのグラフを比較すると、目的に応じてテキストの表し方を変える必要があることがわかる



検出法

- Wikipedia　日本語　約65万記事(4.3GB)のインデクス(suffix array)を事前に作成



- Manber and Myers "Suffix arrays: a new method for on-line string searches". *SIAM Journal on Computing*, Vol. 22, pp. 935-948, 1993.
- SAIS: An implementation of the induced sorting algorithmを利用
- コピペ判定支援ソフトウェア:コピペルナー
  (金沢工業大学　杉光教授考案)

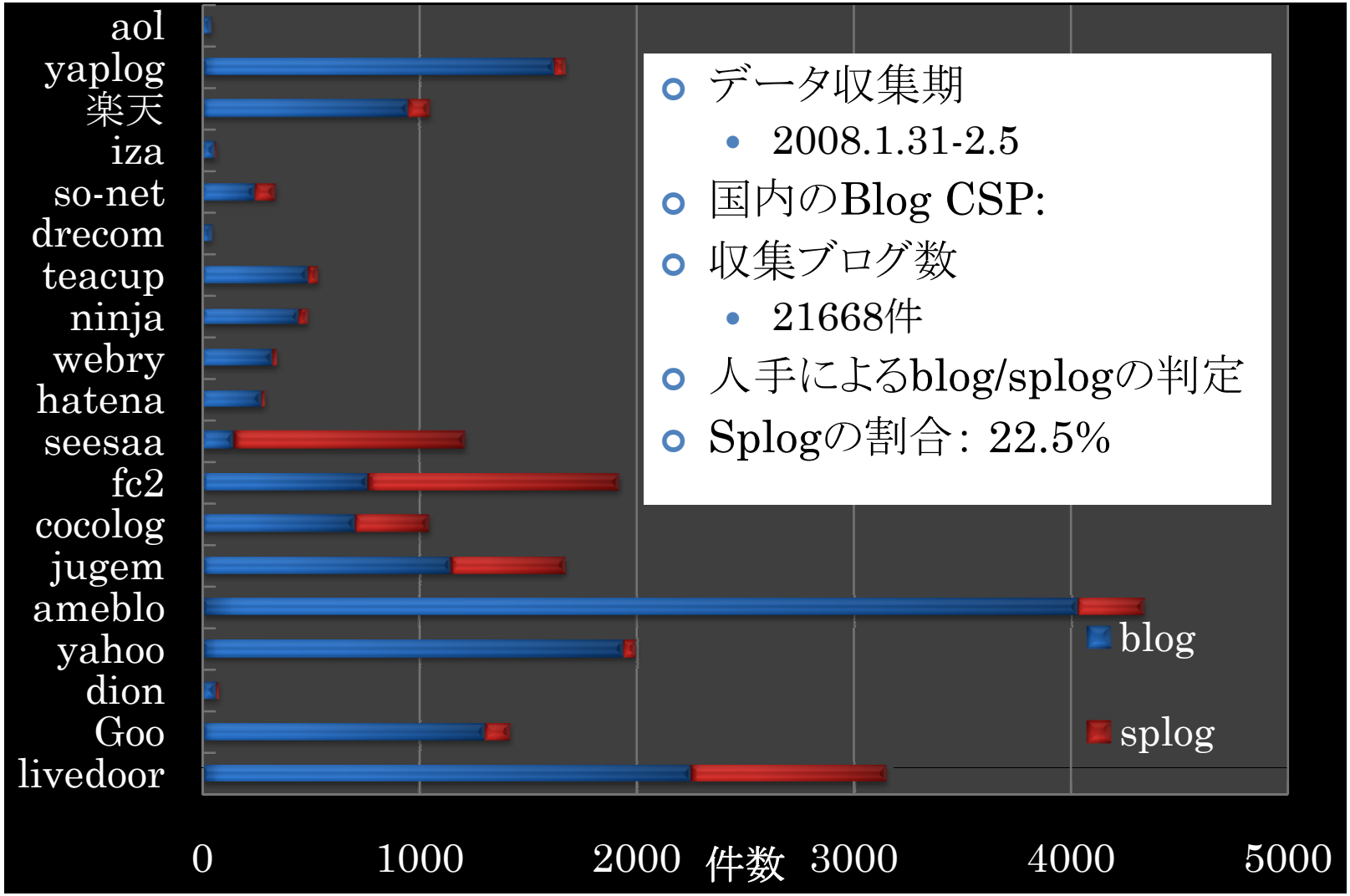
レポート解析

スプログ検出

- スプログ(スパムブログ)： 検索エンジンからのアクセスを増やすために様々なキーワードを埋め込み広告誘導を目的としたブログ。テキストを自動生成したり、他のコンテンツからコピーして生成することが多い
- IICPの調査研究では、2008年1月のスプログの割合は12%程度
- スプログの生成法
  - ワードサラダ型：単語やフレーズを自動的に組合わせて生成
  - コピー＆ペースト型：他のコンテンツ（webニュース、メールマガジン、辞典、QAサイト、RSS　等）
  - テンプレート型：「今日は・・・でした」といったテンプレートにキーワードを埋めこむ
  - Web検索型：流行のキーワードで検索して、上位ページをコピー

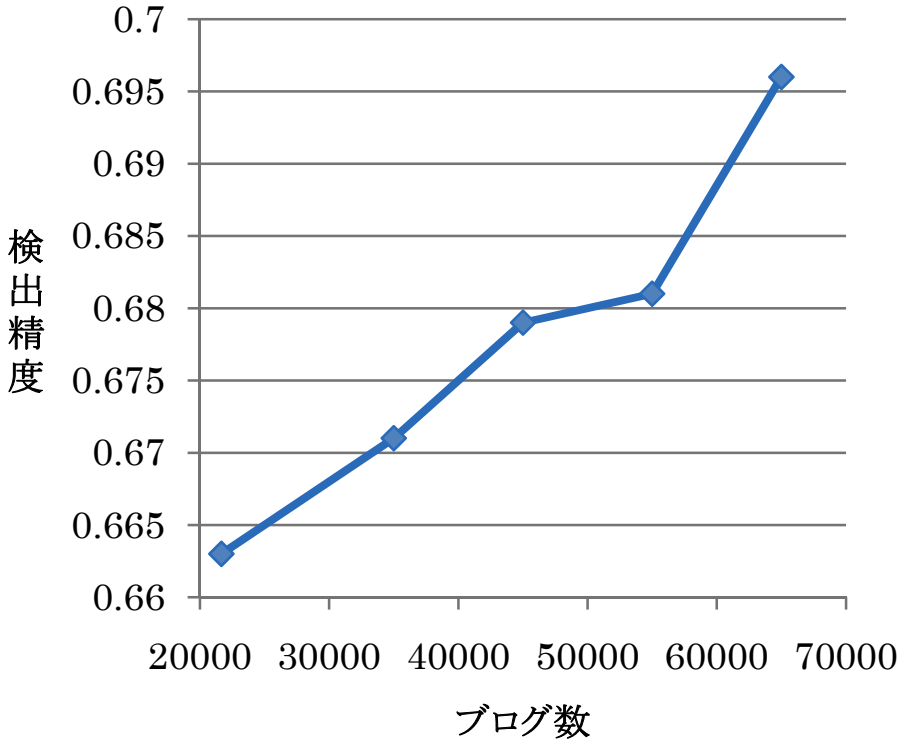
竹田隆治:
“局所的類似情報に基づいたテキストマイニングに関する研究”
総研大　博士論文, 2009　より

スプログの割合


aol
yaplog
楽天
iza
so-net
drecom
teacup
ninja
webry
hatena
seesaa
fc2
cocolog
jugem
ameblo
yahoo
dion
Goo
livedoor
データ収集期
2008.1.31-2.5
国内のBlog CSP:
収集ブログ数
21668件
人手によるblog/splogの判定
Splogの割合：22.5%
blog
splog
0
1000
2000
件数
3000
4000
5000

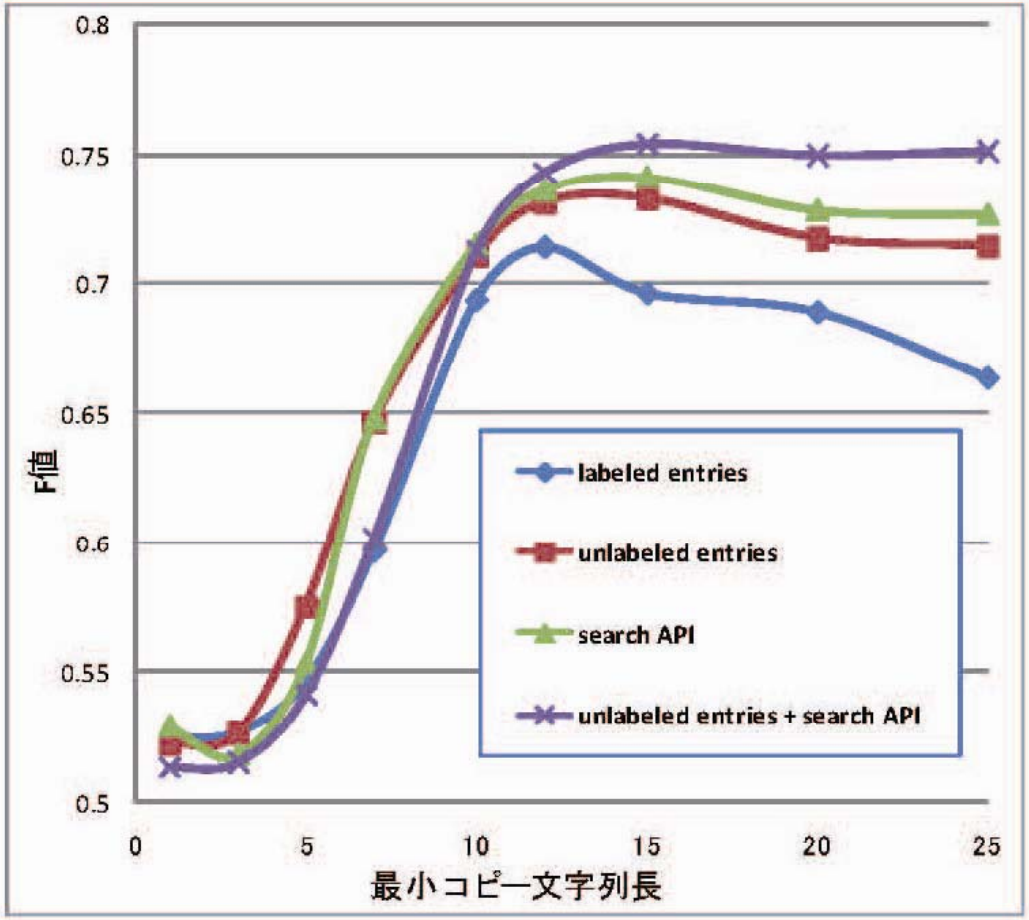
コピー検出に使用するブログ数と検出精度

- スプログは基本的にコピー
- コピーのもととなるテキストをできるだけたくさん用意し、コピー部分を見つけることでsplog検出精度の向上をめざす



文字列マッチングによるコピー検出の課題

- 最小コピー文字列長の設定にノウハウが必要
- 短い単語を複数のサイトからコピーした場合は、検出が難しい



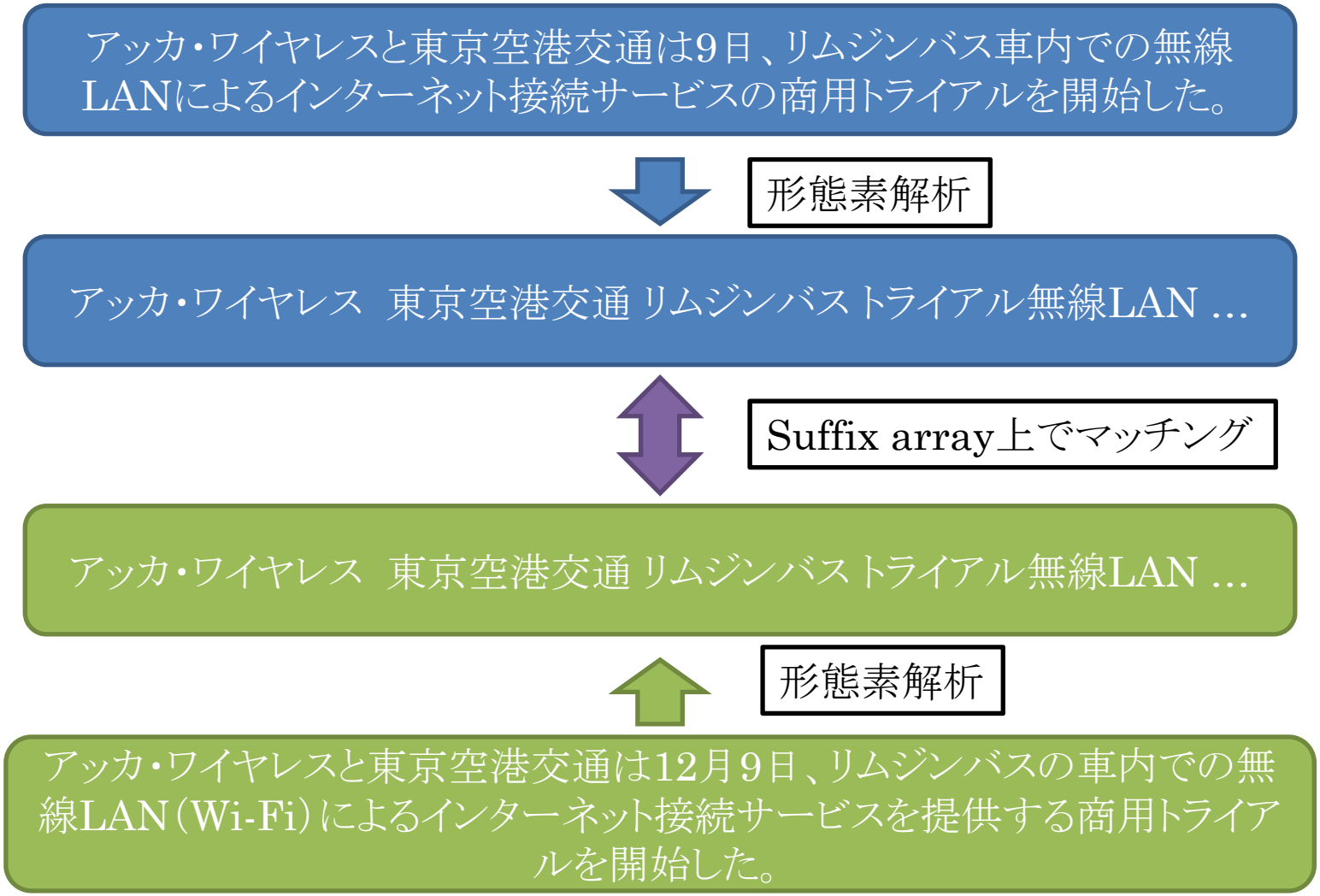
文書要約への応用

アッカ・ワイヤレスと東京空港交通は12月9日、リムジンバスの車内での無線LAN（Wi-Fi）によるインターネット接続サービスを提供する商用トライアルを開始した。成田空港開港30周年を記念し、今年5月から成田空港路線を中心に運行しているラッピングバス「リムちゃん号」の2車両でトライアルを実施。 今後は同一ネットワークを活用して、リムジンバスの車内でディスプレイなどを利用して広告を配信するデジタルサイネージによる地域情報やCM掲示などの整備も検討していく。
両社では、トライアル期間を通じて、ユーザーの利便性などを検証した上で、有料サービスへの移行を検討する。HSDPA網はソフトバンクモバイルの回線を利用し、下りの最大通信速度は3.6Mbps。また、車内に設置されたモバイルルータは、IEEE 802.11b/gに準拠する。トライアルは2009年4月 上旬までを予定し、期間中の利用料金は無料。

多文書要約

- 同じトピックについて記述された複数の文書をまとめて、要約を作成する問題
- おもな手順
  - 同一のトピックを扱う文書の収集
  - 記述内容が重複する部分の除去
  - 重要な内容を含んだ箇所を抽出
- この研究の着目点
  - 重複コンテンツをみつけることで
    - 同一のトピックを扱う文書を収集
    - 重要な文の選択

重要度順に並べた単語列で比較し文のグループ化



要約の生成

- 同一トピックに関する文書
  - 同じグループに属する文を共有している
- 文の重要度
  - 多くの文書に共有されている文
  - 元の文書での位置
  - 含まれる単語の重要性

要約文書の構成

アッカ・ワイヤレスと東京空港交通は12月9日、リムジンバスの車内での無線LAN（Wi-Fi）によるインターネット接続サービスを提供する商用トライアルを開始した。成田空港開港30周年を記念し、今年5月から成田空港路線を中心に運行しているラッピングバス「リムちゃん号」の2車両でトライアルを実施。今後は同一ネットワークを活用して、リムジンバスの車内でディスプレイなどを利用して広告を配信するデジタルサイネージによる地域情報やCM掲示などの整備も検討していく。
両社では、トライアル期間を通じて、ユーザーの利便性などを検証した上で、有料サービスへの移行を検討する。HSDPA網はソフトバンクモバイルの回線を利用し、下りの最大通信速度は3.6Mbps。また、車内に設置されたモバイルルータは、IEEE 802.11b/gに準拠する。トライアルは2009年4月 上旬までを予定し、期間中の利用料金は無料。

ITMedia 2008.12.9

無線LAN対応の携帯音楽プレーヤーやゲーム機、PCなどを所有する利用者がラッピングバスに乗車した際、車内で高速インターネットを利用できるようになる。

ITMedia 2008.12.9　より

Impress 2008.12.9　より

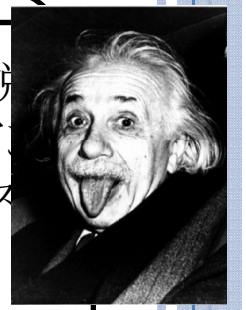
ニュース記事におけるトピックトラッキング

- 重要語に基づいた単語並び替え文の利用例
  - 情報源の異なる記事でも、同じ重要語が文に含まれることが多い
  - この性質を使って、同じ事件について複数の新聞記事をある程度の精度でまとめることができる
- ニュース記事クラスタリングの例
  - 2008年度のノーベル賞
  - 相撲協会のニュース
  - 宮城地震

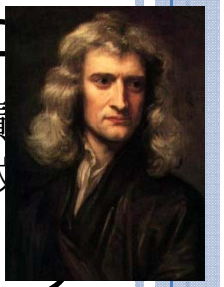
潜在トピックの抽出

潜在トピックとは(1/2)

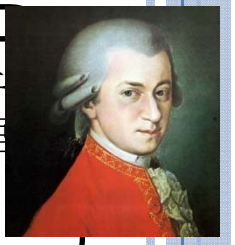
特殊相対性理論及び一般相対性理論、相対性宇宙論、ブラウン運動の起源を説明する揺動散逸定理、光子仮説による光の粒子と波動の二重性、アインシュタインの固体比熱理論、零点エネルギー、半古典型のシュレディンガー方程式、ボース＝アインシュタイン凝縮などを提唱した業績により、20世紀最大の物理学者とも、現代物理学の父とも呼ばれる。



（古典）力学を確立し近代物理学の祖となる。また数学において極めて大きな業績を残した。古典力学は自然科学・工学・技術の分野の基礎となるもので近代科学文明の設立に与えたその影響は計り知れない。



オーストリアの作曲家、演奏家。古典派音楽の代表であり、ハイドン、ベートーヴェンと並んでウィーン古典派三大巨匠の一人である。称号は神聖ローマ帝国皇室宮廷室内作曲家、神聖ローマ帝国皇室クラヴィーア教師、ヴェローナのアカデミア・フィラルモニカ名誉楽長などを勤めた。



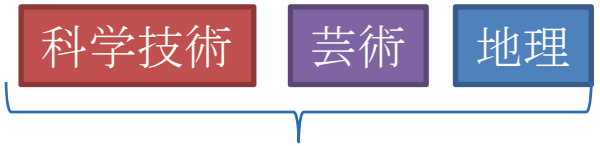
Wikipediaより

潜在トピックとは(2/2)

特殊相対性理論及び一般相対性理論、相対性宇宙論、ブラウン運動の起源を説明する揺動散逸定理、光子仮説による光の粒子と波動の二重性、アインシュタインの固体比熱理論、零点エネルギー、半古典型のシュレディンガー方程式、ボーズ＝アインシュタイン凝縮などを提唱した業績により、20世紀最大の物理学者とも、現代物理学の父とも呼ばれる。

(古典)力学を確立し近代物理学の祖となる。また数学において極めて大きな業績を残した。古典力学は自然科学・工学・技術の分野の基礎となるもので近代科学文明の設立に与えたその影響は計り知れない。

オーストリアの作曲家、演奏家。古典派音楽の代表であり、ハイドン、ベートーヴェンと並んでウィーン古典派三大巨匠の一人である。称号は神聖ローマ帝国皇室宮廷室内作曲家、神聖ローマ帝国皇室クラヴィーア教師、ヴェローナのアカデミア・フィラルモニカ名誉楽長などを勤めた。



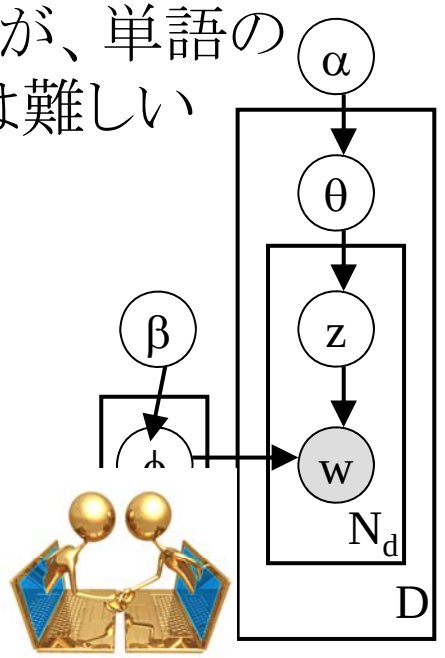
情報統合

- 同じ種類の情報が異なる場所に(異なる形式で)存在する
  - レストランの情報
  - 人の情報(e.g. 年金問題 , spysee)
  - 商品情報(メーカーの商品一覧、消費者レビュー)
  - ホテルの情報
  - ・・・・
- 情報統合は、これらの分散した情報をまとめて提示するための技術
- 世界の研究者との連携の枠組み
  - WEPS: Searching Information about Entities in the Web

人物の識別問題

- 人物のマッチングは情報統合の代表的な問題
  - 同姓同名の人物が同一人物であるかどうかを判定する
  - (似ているが) 異なった表記の人物が同一人物であるかどうかを判定する問題
- 人物に付随する情報から判断することになるが、単語の一致や文字列の一致だけから判断することは難しい

- トピックを使って統合の精度をはかる
- トピックと語の関係はWebの情報を利用する



トピック抽出の手順

特殊相対性理論及び一般相対性理論、相対性宇宙論、ブラウン運動の起源を説明する揺動散逸定理、光子仮説による・・・

形態素解析

特殊 / 相対性理論 / 及び / 一般 / 相対性理論 / 相対 / 性 / 宇宙 / 論 / ブラウン / 運動 / の / 起源 / を / 説明 / する / 揺動 / 散逸 / 定理 / 光子 / 仮説 / に / よる / ・・・

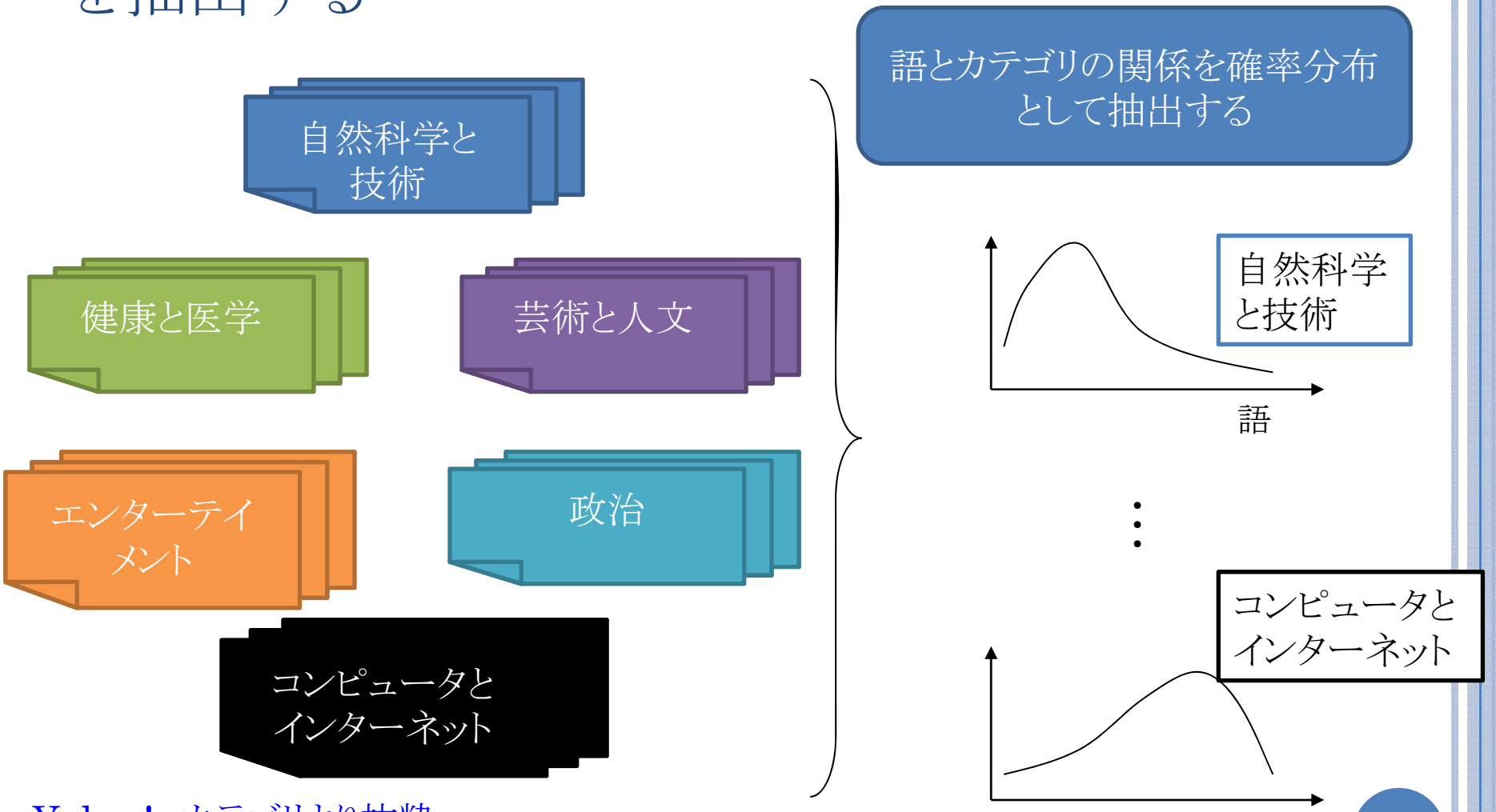
数万～数十万の異なり語よりなるBag-of-Word

トピック割当

特殊 / 相対性理論 / 及び / 一般 / 相対性理論 / 相対 / 性 / 宇宙 / 論 / ブラウン / 運動 / の / 起源 / を / 説明 / する / 揺動 / 散逸 / 定理 / 光子 / 仮説 / に / よる / ・・・

数10～数百のトピックよりなるBag-of-Word

WEBのディレクトリを使って語とトピックの関係を抽出する



Yahoo! カテゴリより抜粋

数万語のBag-of-words → 数百語のトピック

人物に関する記述からトピックを推定し判定

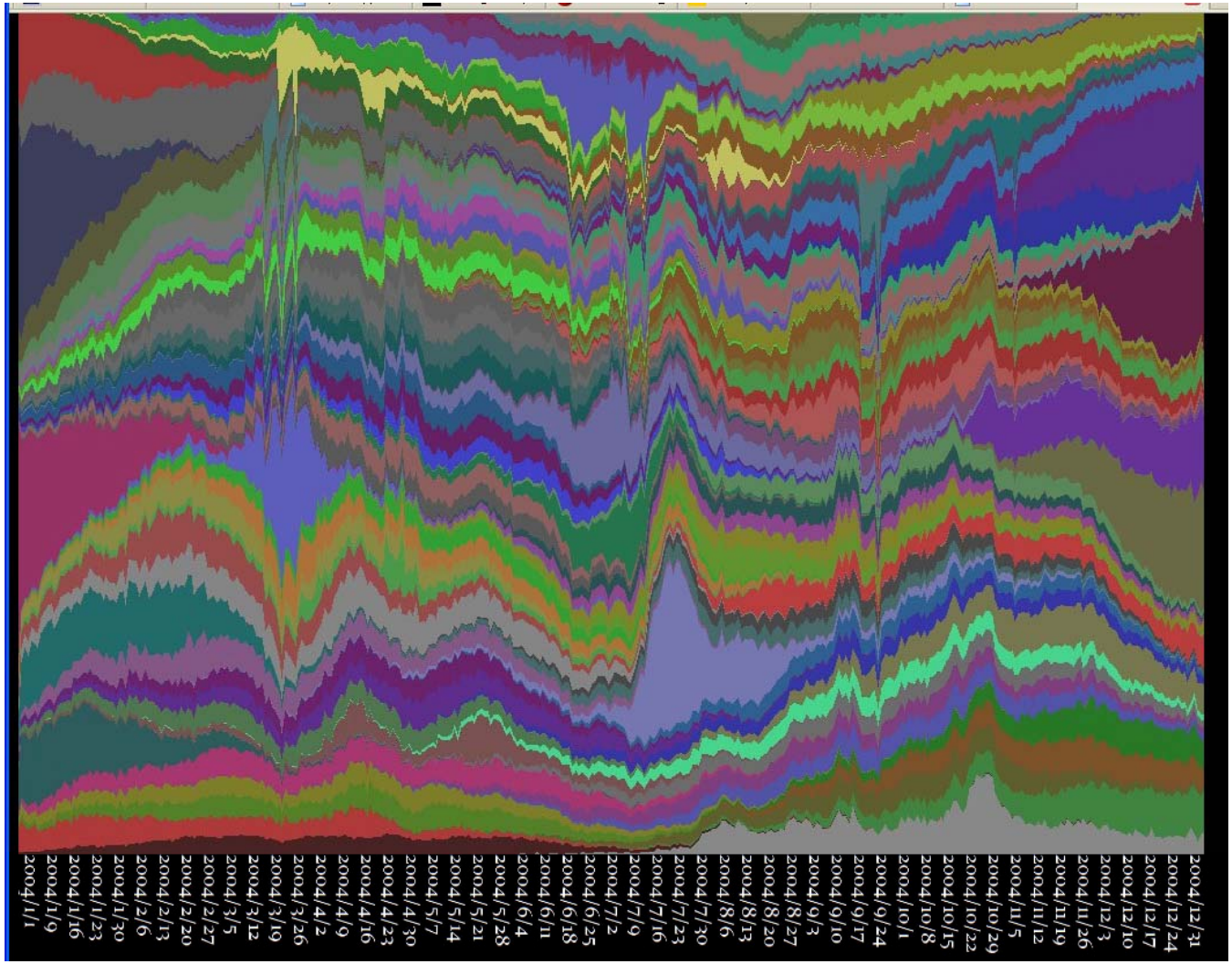
- 人名でWeb検索を行い人物に関する記述から、その人物のトピックを推定
- トピックの類似する人物は同一の人物と判定する
- トピックを使用しない場合と比較して、およそ7%程度精度向上

氏名	固有名抽出	ベクタ空間	Minh08
Sakai	73.8	77.1	**79.7**
Mitchell	71.9	79.9	**81.5**
Tanaka	**79.5**	68.6	72.4
Lafferty	76.9	81.4	**89.7**
Taylor	74.6	74.1	**85.6**
Bridge	**55.8**	50.9	55.1
Ford	52.6	51.9	**72.0**
Hammer	76.2	68.6	**82.0**
Patterson	57.1	65.0	**81.5**
Hooper	44.3	52.5	**60.0**
Roberts	76.9	74.0	**81.1**
Woods	**91.4**	90.7	84.5
Cleveland	60.4	59.0	**80.8**
Average	67.6	68.6	75.1

Q. M. Vu, A. Takasu, J. Adachi:
“ Name Disambiguation Boosted by Latent Topics from Web Directories, ”
IEEE/WIC/ACM Intl. Conf. Web Intelligence, pp. 697-703, 2008 より

ニュース記事からのトピック変遷マップの抽出

テキストストーリ中に現れたトピック数



時間

詳しくはこちらをご覧ください
http://www.cis.nagasaki-u.ac.jp/~masada/researches.html

時系列文書の各時点でのトピックの出現頻度

- 時系列で配送されテキストから関連するトピックを抽出
- 新聞記事からトピック抽出
  - 時間情報を考慮したトピック抽出例 （2005）
    - 抽出されたトピック
      - 1月：スマトラ沖地震、5月：JR西日本脱線事故
      - 9月：郵政選挙、 秋～冬：鳥インフルエンザ東南アジアで流行
      - 11月：耐震偽装 等々
    - 約15万記事（5000万語）の解析
    - 処理時間 16時間
    - 他国のニュース
      - 韓国語ニュース、 中国語ニュース
- 学術論文からのトピック抽出
  - 情報処理分野論文DB（DBLP）

謝辞

- この資料には下記研究者との共同研究の成果が含まれています。ここに感謝いたします。

  - 国立情報学研究所　安達　淳　先生
  - 長崎大学　正田　備也　先生
  - 国立情報学研究所　深川大路　さん
  - サイボーズ　Vu Quang Minh　さん
  - 総研大情報学専攻　竹田隆治　さん