トピックの特性に注目した ブログマイニングとその周辺

筑波大学大学院
システム情報工学研究科 知能機能システム専攻
宇津呂武仁

2010年9月28日(火)、『言語処理技術の深化と理論・応用の新展開』科研・合同シンポジウム、@東大

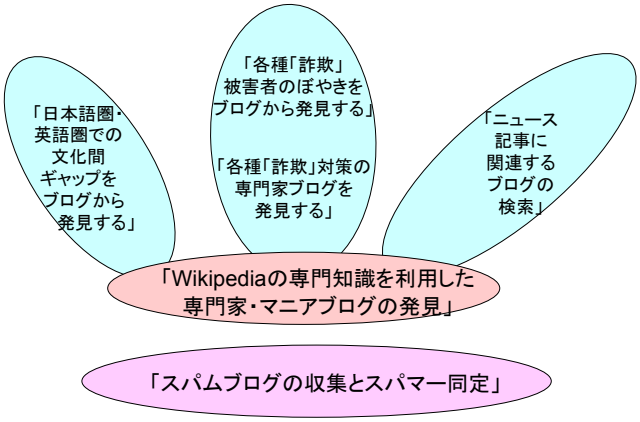
ブログマイニング

- ブログの特性
  - 個人・団体等の時系列な情報提示・日記
  - 書き手の顔(立場・意見・職業)が見える
  - 情報(記述内容)の類型化の戦略が立てやすい?

ブログマイニング:本研究のアプローチ

- ブログの特性
  - 個人・団体等の時系列な情報提示・日記
  - 書き手の顔(立場・意見・職業)が見える
  - 情報(記述内容)の類型化の戦略が立てやすい?
- ブログからマイニングしたい情報
  - 専門家ブロガーの関心事項・意見動向
  - 体験者による臨場感ある描写
  - 異なる立場・国籍・言語のブロガーによる情報の収集
    ⇒比較・対象分析

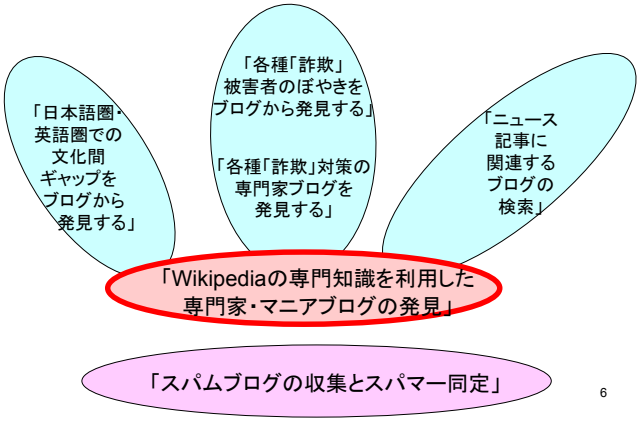
ブログマイニング:研究事例



研究協力体制

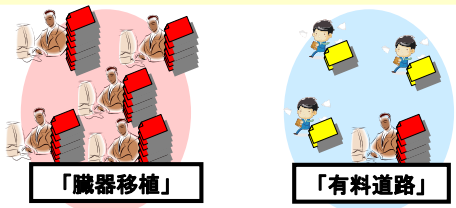
- 科研費・基盤B, 平成20～22年度
- 「トピックの特性を言語間で比較・対照分析する多言語ウェブテキストマイニングの研究」
- NII共同研究(公募研究)、平成19年度～
  「多言語情報に関する戦略的情報アクセスシステム」
- 「世界ニュース」研究グループ
  - 神門(NII)、中川(東大)、吉岡(北大)
  - 清田(東大)、福原(産総研)
  - 宇津呂(筑波大)、他

ブログマイニング:研究事例



研究の目的

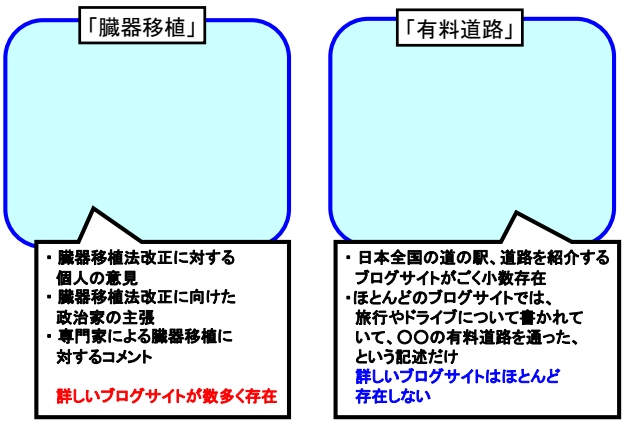
あるトピックについて
・どのくらい詳しいブログサイトが書かれているか
・個々のブログサイトがどのくらい詳しく書かれているか
を自動的に判定する



詳しいブログサイト：多い
個々のブログサイト：非常に詳しい

詳しいブログサイト：
ほとんど存在しない

トピックについて詳しいブログサイトの例



トピックについて、どれだけ詳しいブログが書かれているかを分析した結果

非常に詳しく書かれている	・臓器移植	・アルコール依存症
	・著作権侵害	・オークション詐欺
	・フィッシング詐欺	・園芸
比較的詳しく書かれている	・バイオインフォマティクス	
	・フコイダン	・ダイコン
ほとんど書かれていない	・有料道路	・武田薬品工業
	・トムラウシ山	・劇場
	・リモコン	・関東

目次

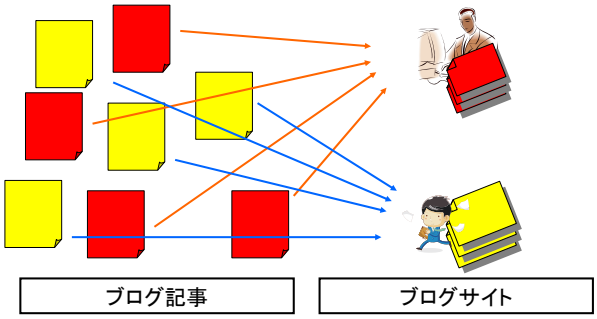
- 研究の目的
- トピックについて詳しいブログの選択的検索
- どのくらい詳しくブログが書かれているトピックなのか分析
- まとめと今後の課題

「臓器移植」について詳しく書かれたブログサイトを選択的に検索

例：「臓器移植」

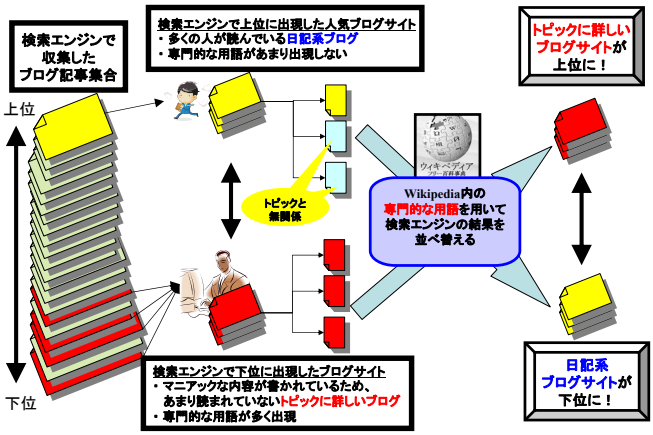
どのくらい詳しいブログが書かれているトピックか分析するためには詳しく書かれたブログを選択的に検索することが必要

検索の対象とするブログの単位



同一著者によるブログ記事のまとまり

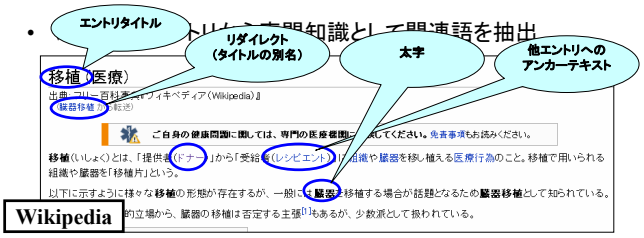
提案手法:トピックに詳しいブログサイトの収集



どのような知識を用いてブログサイトを順位付けするか？

- エントリタイトル
- リダイレクト（タイトルの別名）
- 太字
- 他エントリへのアンカーテキスト

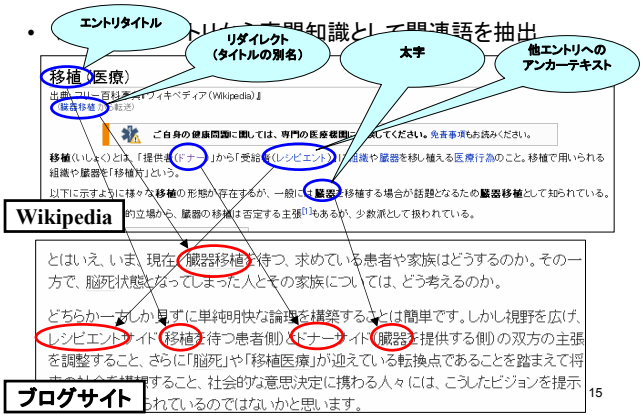
を抽出



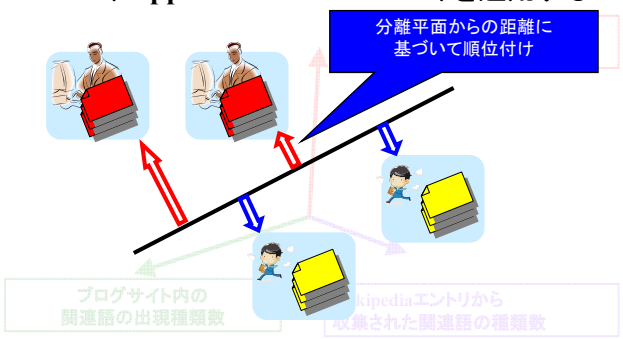
どのような知識を用いてブログサイトを順位付けするか？

- エントリタイトル
- リダイレクト（タイトルの別名）
- 太字
- 他エントリへのアンカーテキスト

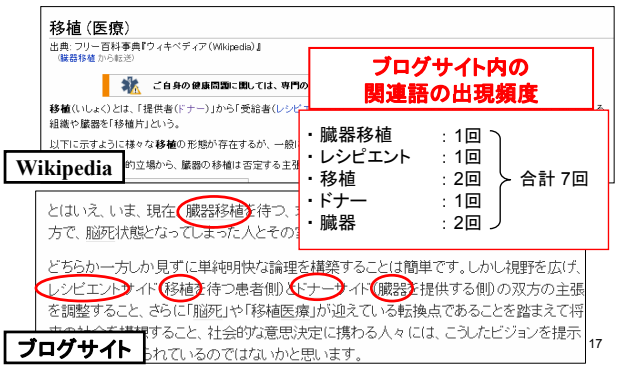
を抽出



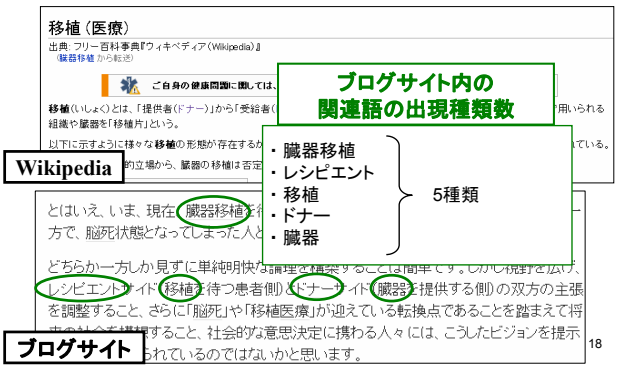
関連語の特徴量を用いて SVM（Support Vector Machines）を適用する



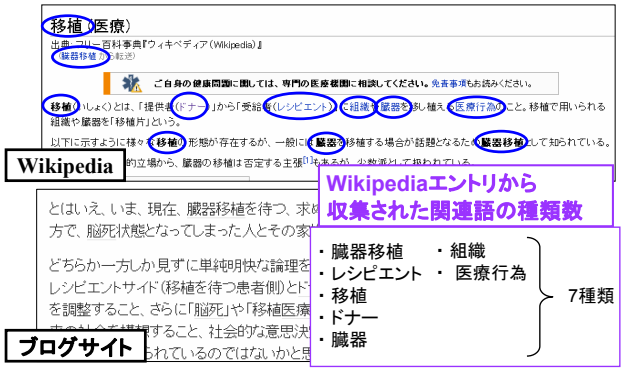
SVMで用いる特徴量



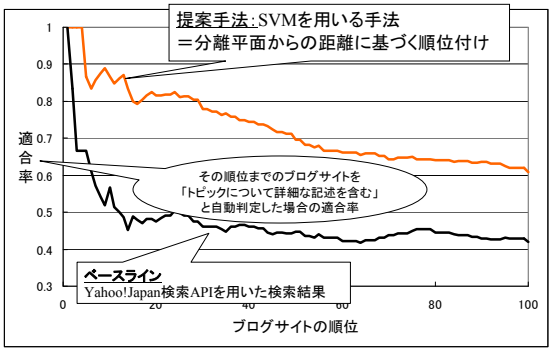
SVMで用いる特徴量



SVMで用いる特徴量

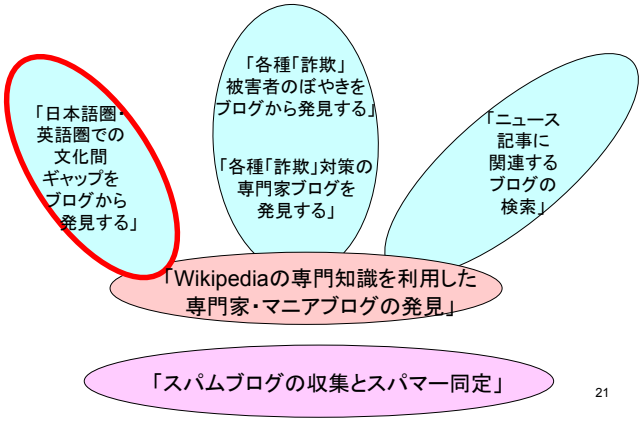


順位付け性能の比較



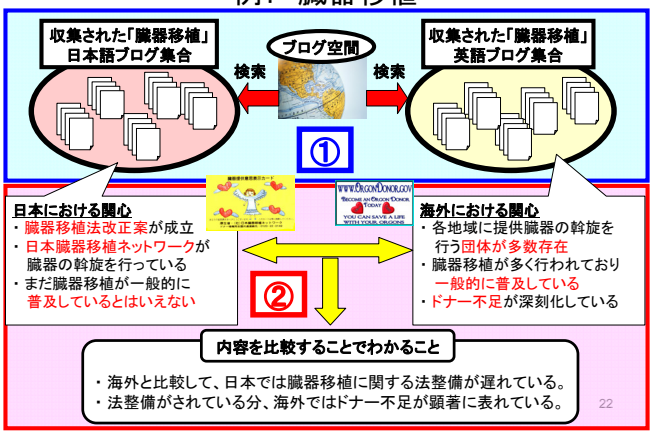
20

ブログマイニング：研究事例



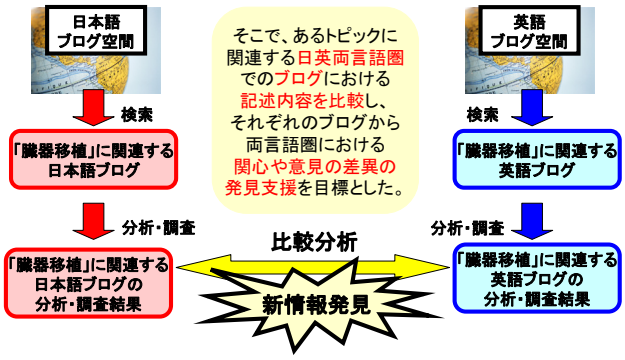
21

例： 臓器移植



22

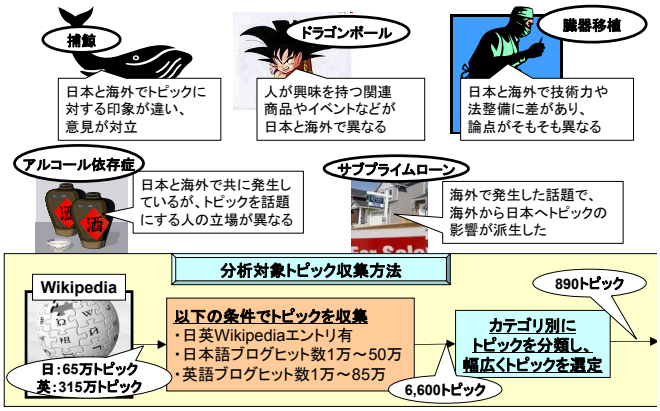
研究の全体的枠組み



23

どういったトピックを比較したか？

- **日英でそれなりに話題になっているが、日英間で内容の差異がありそう**



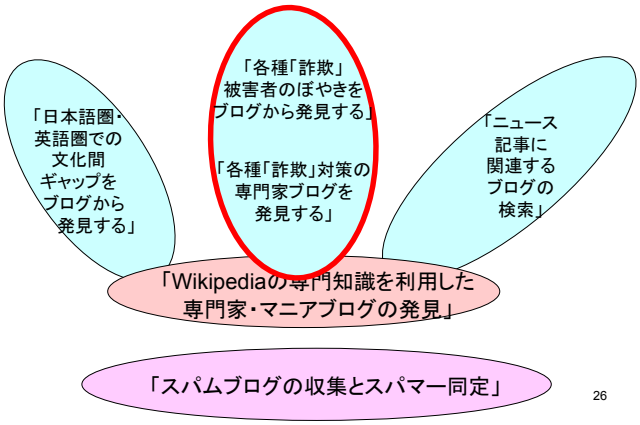
分析済:21トピック

- ブログにおける日英間の差異による定性的な分類

日英間の差異: 大	**5トピック**
日英間の差異: 中	**7トピック**
日英間の差異: 小	**9トピック**

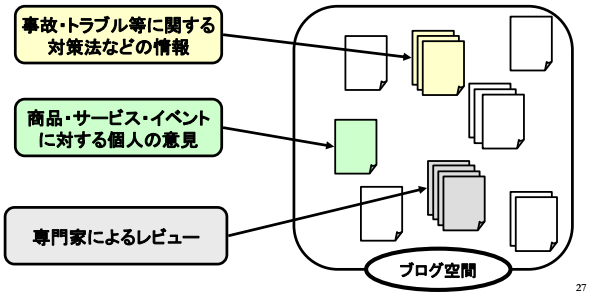
・日英各言語のブログにおける記述の特徴を分析
・各言語のブログの特徴から、日英言語間における内容の差異を分析

ブログマイニング:研究事例



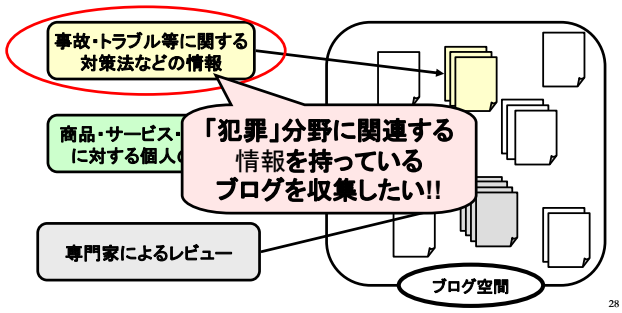
研究の背景

ブログ:個人が自由に意見や情報を発信できるツール
⇒従来の報道からは得られないさまざまな情報が多く存在する。

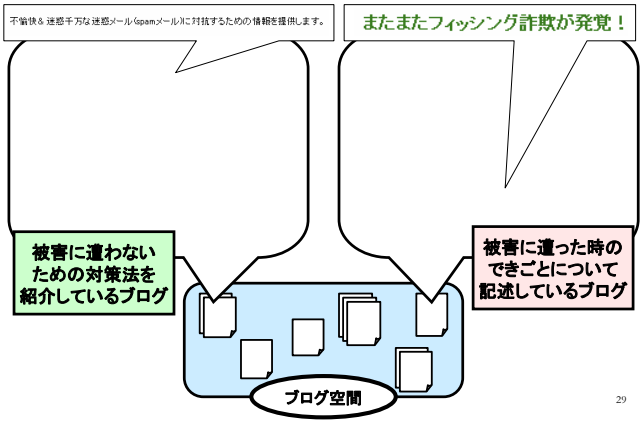


研究の背景

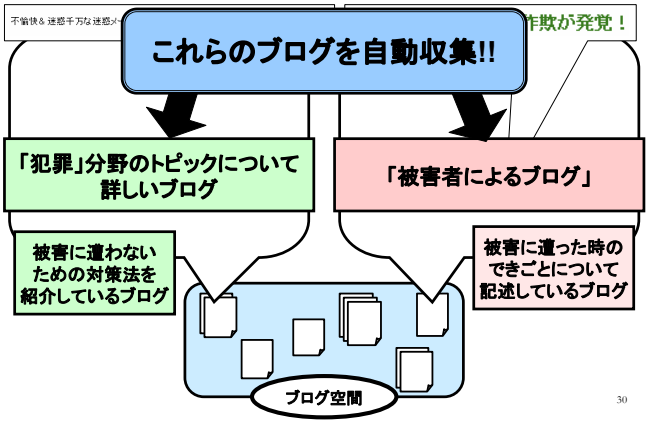
ブログ:個人が自由に意見や情報を発信できるツール
⇒従来の報道からは得られないさまざまな情報が多く存在する。



「犯罪」分野に関連するブログの例



本研究の目的



背景：消費者生活センター

- 国民生活センター・・・国の公的機関
  - 消費者からのトラブルに関する相談事例や対策方法を紹介
  - 消費者が実際に遭遇した幅広い分野のトラブルを公開している
    - ネット・電話、衣食住、趣味・レジャー、金融、個人情報、etc..
    - 約300件の相談事例が公開されている（2010年8月現在）
- 地方自治体にも消費者生活センターが存在
  - 東京都消費生活総合センター
  - 名古屋市消費生活センター

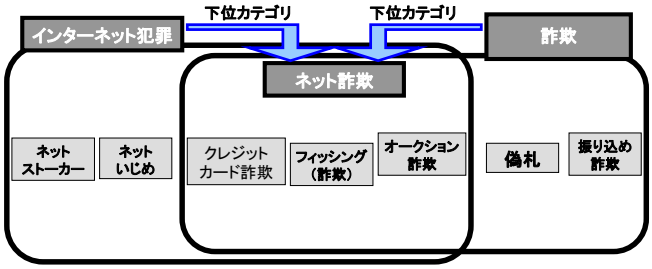
背景：消費者生活センター

- **トラブルに関する対策情報の需要は多い！**
- 消費者からのトラブルに関する相談事例や対策方法を紹介
- 名古屋市消費生活センター

しかし…
・公開されている情報は人手による作業が必要でコストがかかる。
・人手による作業のため、公開されている情報の種類・特徴・数が限られている。

⇒ **幅広い分野における当事者・専門的分析者のより身近な声を自動でブログ空間から多く取得できるか検証！**

今回、対象とした犯罪分野のトピック

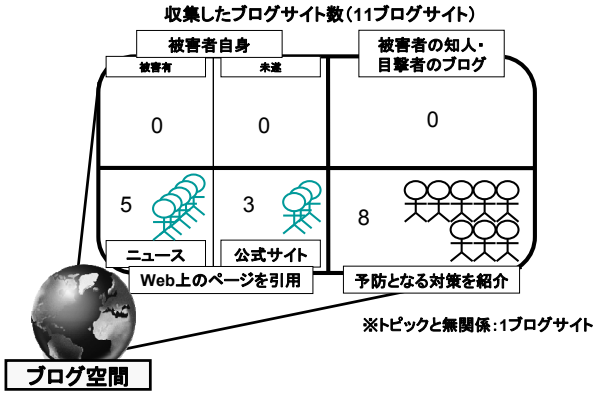


Wikipediaにおけるカテゴリ構造

目次

- 研究の背景・目的
- 「犯罪」分野のトピックについて詳しいブログの収集
- 「被害者によるブログ」の自動収集
- まとめ・今後の課題

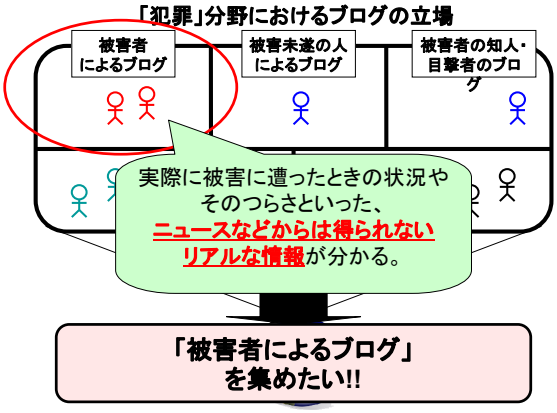
結果：「フィッシング詐欺」



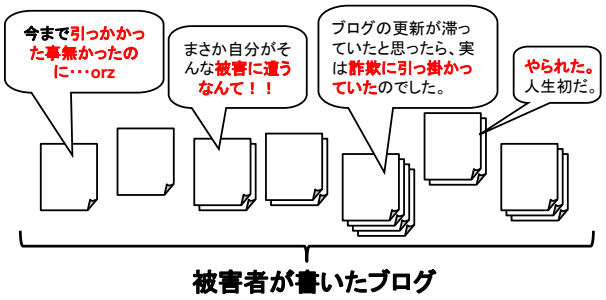
目次

- 研究の背景・目的
- 「犯罪」分野のトピックについて詳しいブログの収集
- 「被害者によるブログ」の自動収集
- まとめ・今後の課題

「被害者によるブログ」に注目する理由

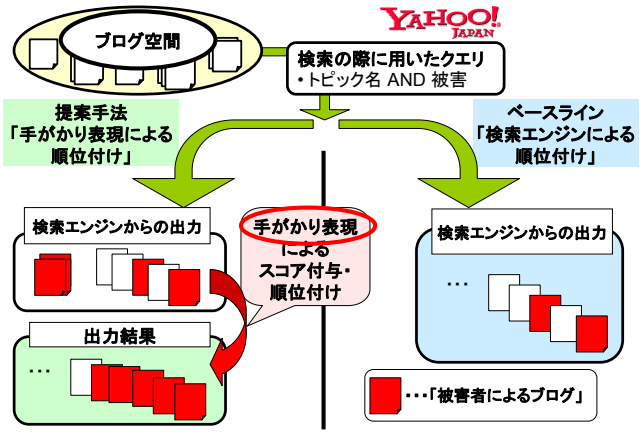


「被害者によるブログ」を見つけるために…



実際に被害者が書いたブログで使われているような、「被害に遭ったと分かる手がかり表現」を集めて、それらを用いて被害者が書いたブログを同定する。

提案手法と検索エンジンとの比較

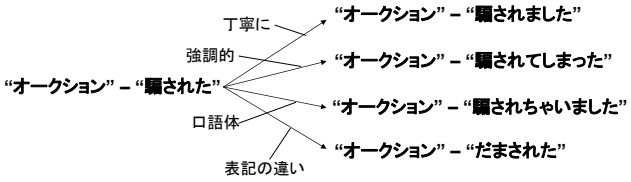


係り受け関係について

例:「オークションで見事に騙された」

→ “オークション” – “騙された”といった、被害を表すと思われる係り受け関係を、手がかり表現とした。

●表現の派生も考慮する



文節単位の表記パターンについて

- 係り受け関係以外に、「被害に遭った」ことを表していそうな言葉も、手がかり表現とした。

例:「やられた」 ⇒ “やられた。人生初だ。”

例:「不審」

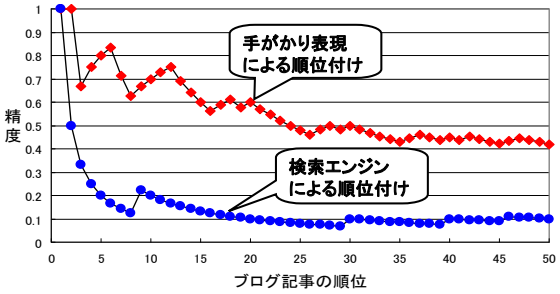
⇒ “○○○カードを使っている私に、「不審な買い物の記録があるんですが・・・」、と切り出され、「え～っ！！」”

手がかり表現一覧

- **手がかり表現に対して、種類に応じた重み**を与えておき、それらを用いて**ブログ記事にスコア付与・順位付け**をおこなった。

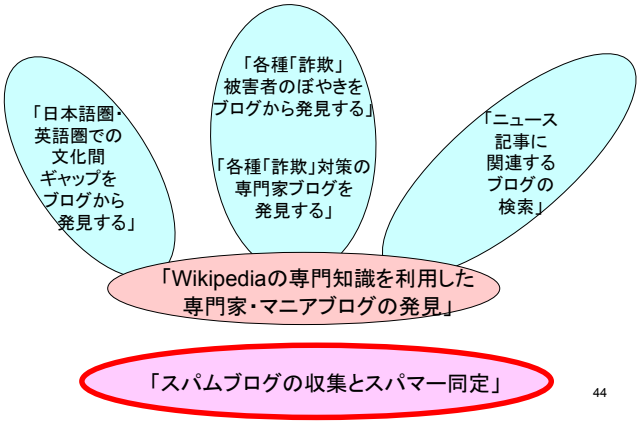
手がかり表現の種類		手がかり表現の例	スコア	種類数			
				活用有り	活用無し	その他	合計
係り受け関係	基本形	被害 – 遭う, 詐欺 – 引っ掛かる	10	13	3	3	19
	派生形	被害 – あいました, 詐欺 – 引っ掛かった		84			90
文節単位の表記パターン	高スコア	やられた, 騙された, 詐欺られた	2	11	1	1	13
	中スコア	音信不通, 凹む, 被害届, 不審	1	15	91	7	113
	低スコア	警察, 連絡, あって	0.5	5	13	0	18

比較結果 (オークション詐欺)



提案手法のほうが、ベースラインよりも
「被害者によるブログ」の検索性能が高い

ブログマイニング：研究事例



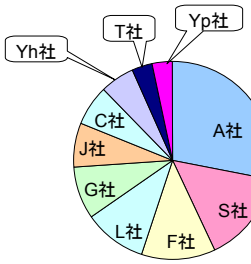
発表の内容

- スプログのHTML構造に着目
- 同一作成者によって作成されたスプログ（大量生成型スプログ）は、HTML構造が類似しているだろうか？
- 主要ブログホスト10社を対象として検証

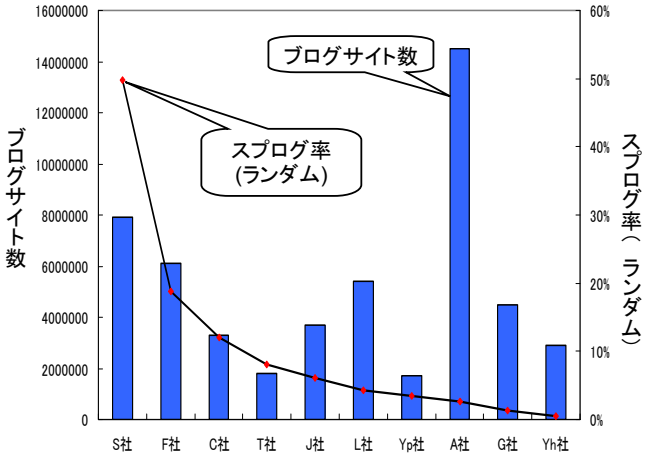
日本語ブログ空間

- 2004年3月から2009年5月までのRSSを取得し、約6万ドメインに対して約620万のブログサイトURLを取得
- 主要ブログホスト会社10社：約520万URL

主要10ホスト分布



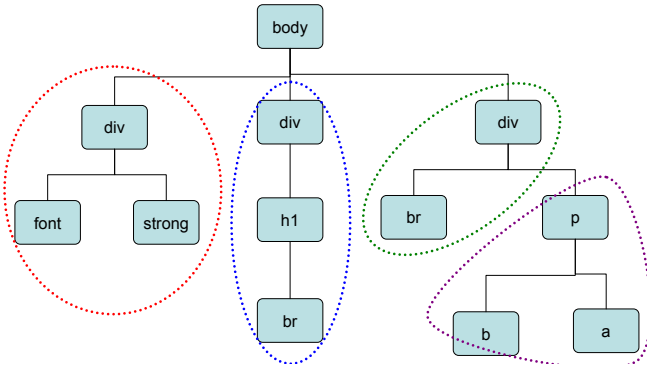
ブログホスト会社	A社	S社	F社	L社	G社	J社	C社	Yh社	T社	Yp社
ブログ数	145万 (28%)	79万 (15%)	61万 (12%)	54万 (10%)	45万 (9%)	37万 (7%)	33万 (6%)	29万 (6%)	18万 (4%)	17万 (3%)



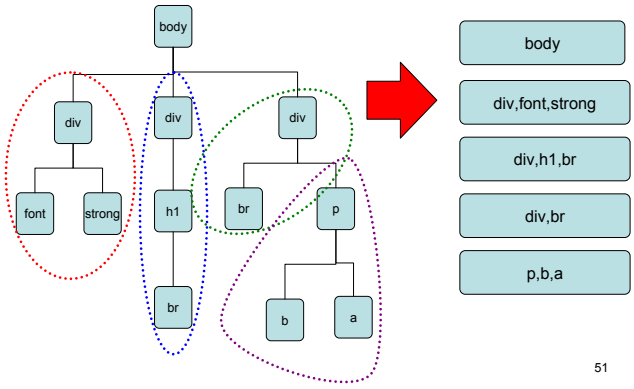
データセット

- 520万URLのうち10分の1の52万URLを収集
- ランダムサンプリングにより
  各ブログホスト会社：500サイト
- それぞれスプログ、非スプログのラベルづけ
  – 同じコメントが存在するものは大量生成型スプログ

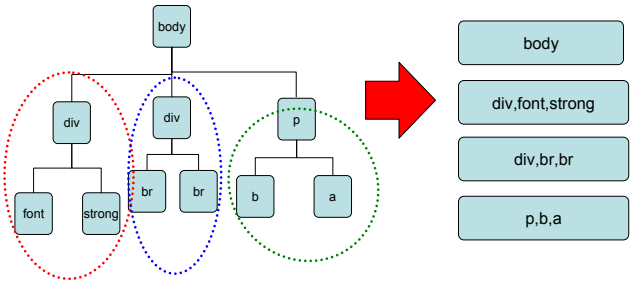
HTMLの木構造をブロックに まとめあげる(ブログサイトs)



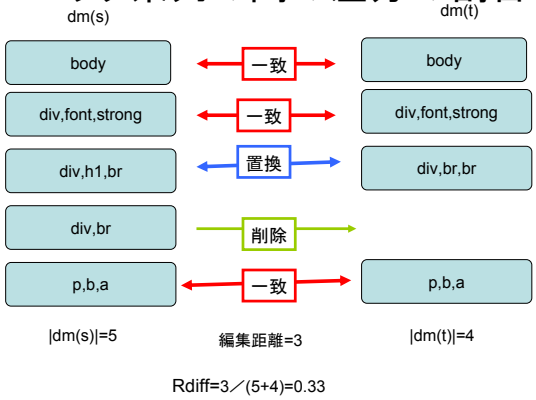
ブロック系列の抽出(ブログサイトs)



ブロック系列の抽出(ブログサイトt)

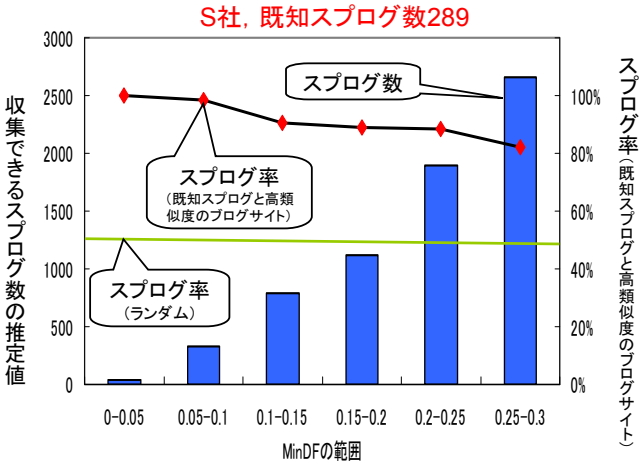


ブロック系列の間の差分の割合

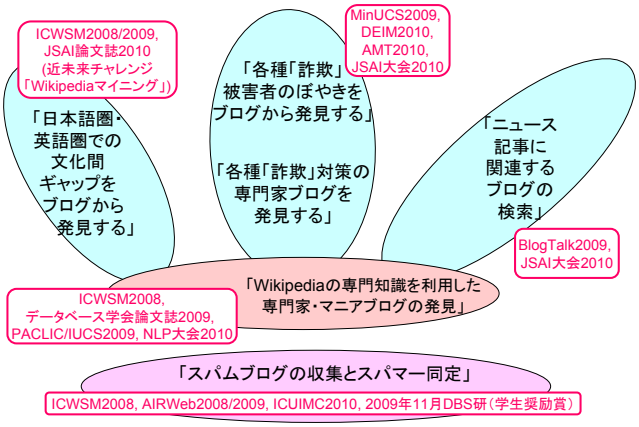


|dm(s)|=5　　編集距離=3　　|dm(t)|=4

Rdiff=3／(5+4)=0.33



ブログマイニング：研究発表



現在取り組んでいる課題

- 専門家ブロガー検索の高精度化
- ブロガー・ブログ記事の自動類型化
  - 関心事項・話題の細分化
  - 時系列素性の導入
- 日英ブログ間の文化間差異発見（支援）
  - 差異の定量化・自動検出
- スパムブログの収集とスパマー同定
  - アフィリエイト情報抽出パターンの効率的獲得
  - アフィリエイトリンクの類型化と機械学習での評価