

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?

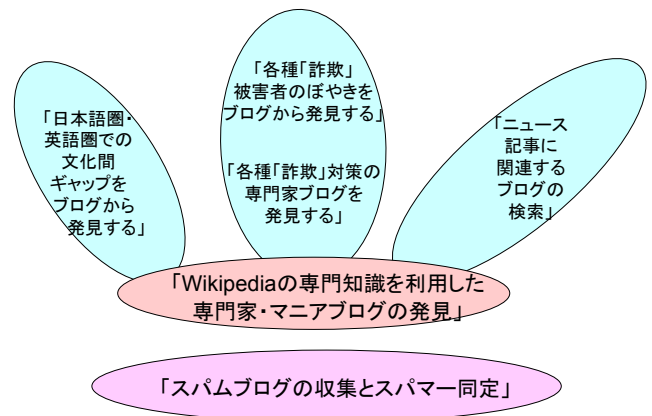
2

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

3

ブログマイニング: 研究事例

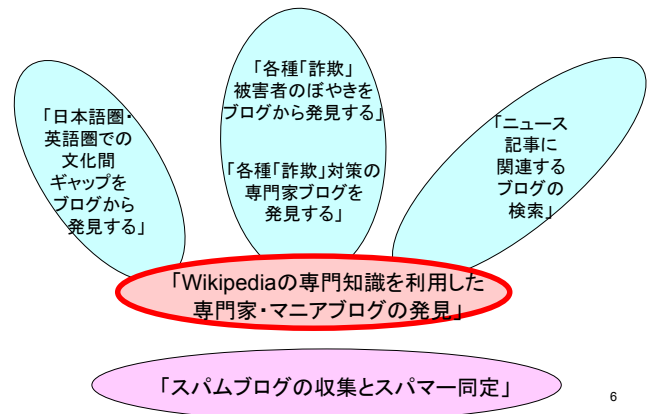


研究協力体制

- 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

5

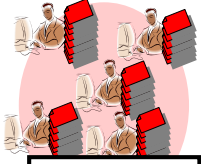
ブログマイニング: 研究事例



6

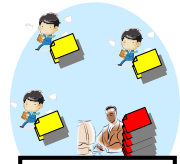
研究の目的

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



「臓器移植」

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



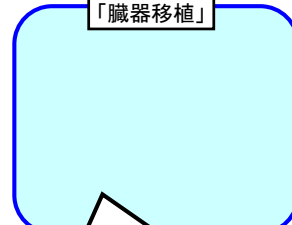
「有料道路」

詳しいブログサイト: ほとんど存在しない

7

トピックについて詳しいブログサイトの例

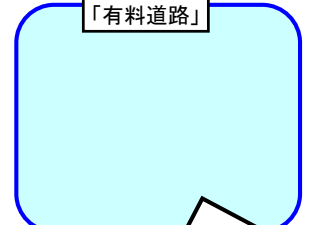
「臓器移植」



- ・臓器移植法改正に対する個人の意見
- ・臓器移植法改正に向けた政治家の主張
- ・専門家による臓器移植に対するコメント

詳しいブログサイトが数多く存在

「有料道路」



- ・日本全国の道の駅、道路を紹介するブログサイトがごく少数存在
- ・ほとんどのブログサイトでは、旅行やドライブについて書かれていて、〇〇の有料道路を通った、という記述だけ

詳しいブログサイトはほとんど存在しない

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
- ・著作権侵害
- ・フィッシング詐欺
- ・アルコール依存症
- ・オークション詐欺
- ・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
- ・トムラウシ山
- ・リモコン
- ・武田薬品工業
- ・劇場
- ・関東

9

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

10

「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」



Web空間

検索

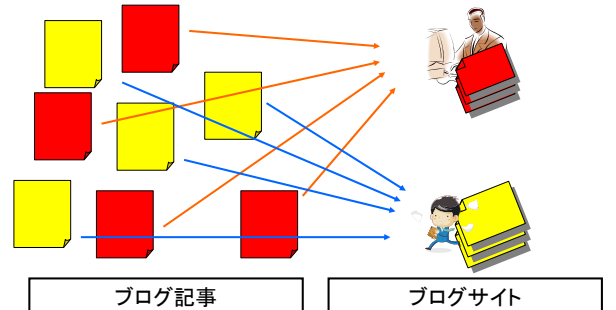
「臓器移植」について詳しく書いているブロガー

「臓器移植」にあまり詳しくないブロガー



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



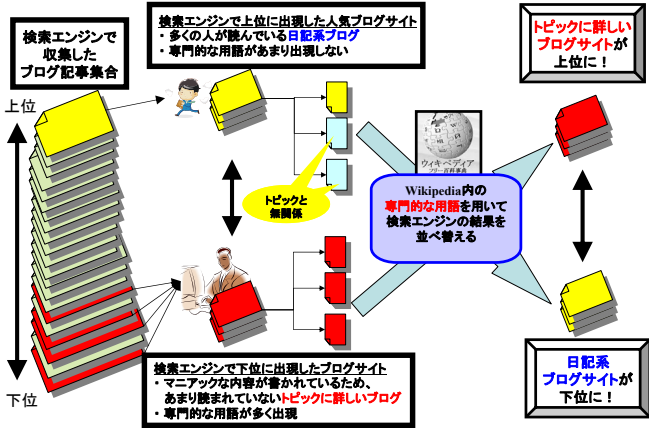
ブログ記事

ブログサイト

同一著者によるブログ記事のまとめ

12

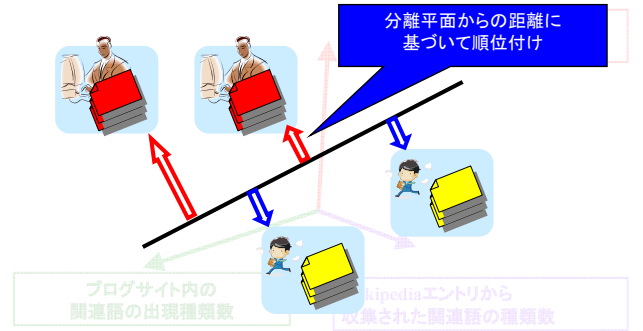
提案手法:トピックに詳しいブログサイトの収集



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?

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?

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

臓器移植	: 1回	} 合計 7回
レシピエント	: 1回	
移植	: 2回	
ドナー	: 1回	
臓器	: 2回	

SVMで用いる特徴量

臓器移植	} 5種類
レシピエント	
移植	
ドナー	
臓器	

SVMで用いる特徴量

移植 (医療)
 出典: 日本百科事典『ウィキペディア (Wikipedia)』
 (後略)

ご自身の健康問題に関しては、専門の医療機関に相談してください。患者さまもお読みください。

移植 (しよく)とは、「提供者 (ドナー) から「受給者 (レシピエント) に組織 (臓器) を移植 (しよく) すること。移植で用いられる組織や臓器を「移植片」という。

以下に示すように様々な「移植」の形態が存在するが、一般に「臓器」を移植する場合は話題となる「臓器移植」して知られている。この立場から、臓器の移植は否定する主張¹⁾もあるが、少数派として扱われている。

Wikipedia

Wikipedia エントリから
 収集された関連語の種類数

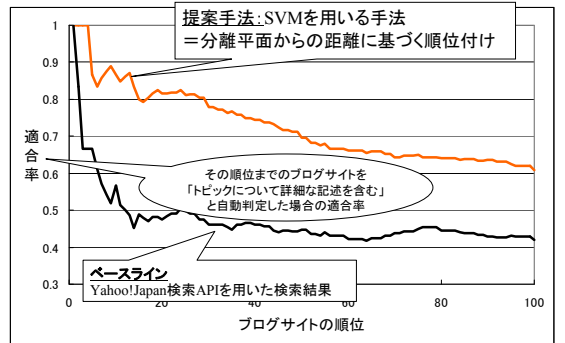
- 臓器移植
- 組織
- レシピエント
- 医療行為
- 移植
- ドナー
- 臓器

7種類

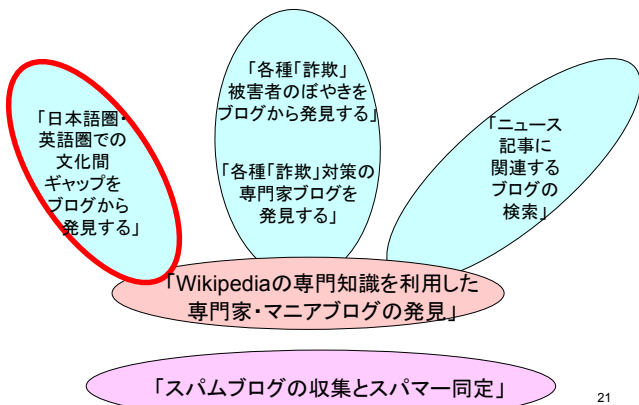
ブログサイト

と「はい、いま、現在、臓器移植を待つ、求む、方、脳死状態となつてしまった人とその家、どちらか一方しか見ずに単純明快な論理をレシピエントサイド (移植を待つ患者側) とを調整すること、さらに「脳死」や「移植医療」を調整すること、社会的な意思決定されているのではないかと

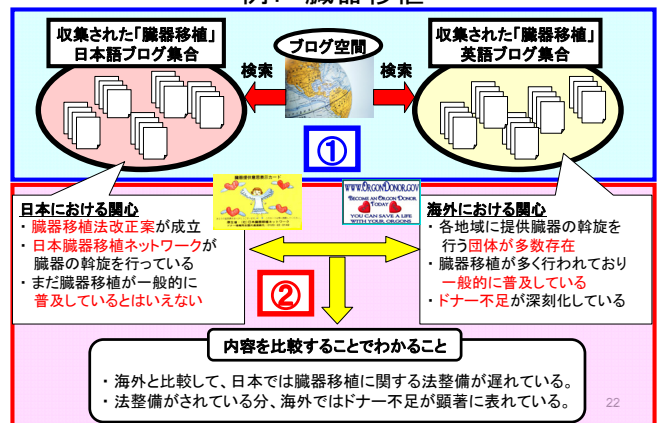
順位付け性能の比較



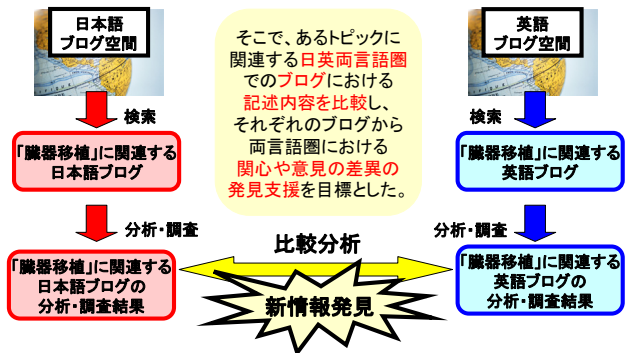
ブログマイニング: 研究事例



例: 臓器移植



研究の全体的枠組み



どういったトピックを比較したか?

日英でそれぞれに話題になっているが、日英間で内容の差異がありそう

- 捕鯨**: 日本と海外でトピックに対する印象が違い、意見が対立
- ドラゴンボール**: 人が興味を持つ関連商品やイベントなどが日本と海外で異なる
- 臓器移植**: 日本と海外で技術力や法整備に差があり、論点がそもそも異なる
- アルコール依存症**: 日本と海外で共に発生しているが、トピックを話題にする人の立場が異なる
- サブプライムローン**: 海外で発生した話題で、海外から日本へトピックの影響が派生した

分析対象トピック収集方法

Wikipedia
 日: 65万トピック
 英: 315万トピック

以下の条件でトピックを収集
 ・日英Wikipediaエントリ有
 ・日本語ブログヒット数1万~50万
 ・英語ブログヒット数1万~85万

890トピック

6,600トピック

カテゴリ別にトピックを分類し、幅広くトピックを選定

分析済:21トピック

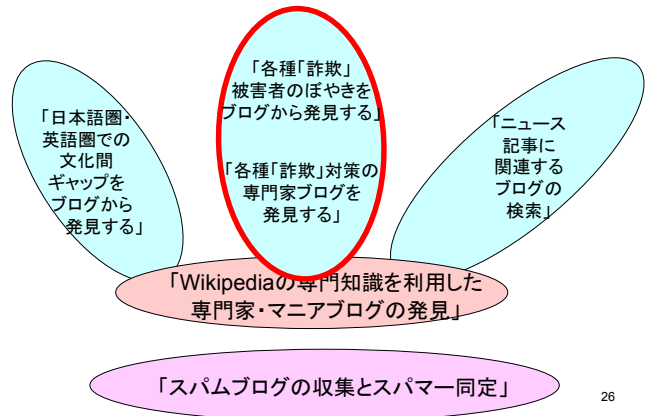
- ブログにおける日英間の差異による定性的な分類

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

25

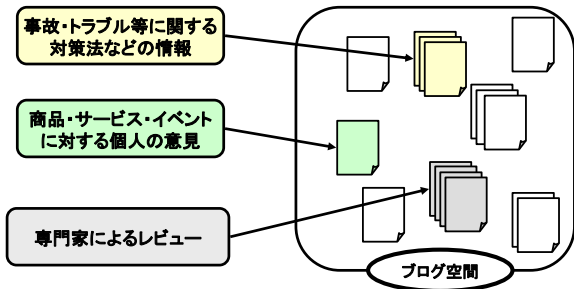
ブログマイニング:研究事例



26

研究の背景

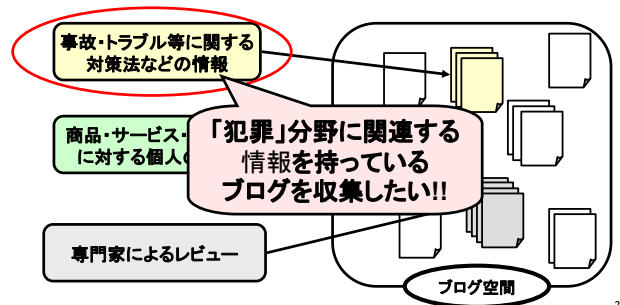
ブログ:個人が自由に意見や情報を発信できるツール
⇒従来の報道からは得られないさまざまな情報が多く存在する。



27

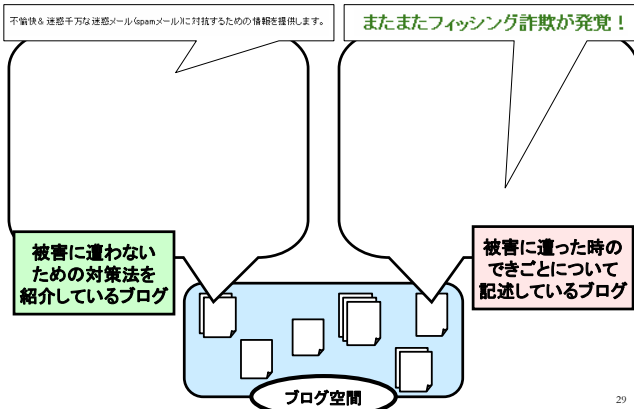
研究の背景

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。



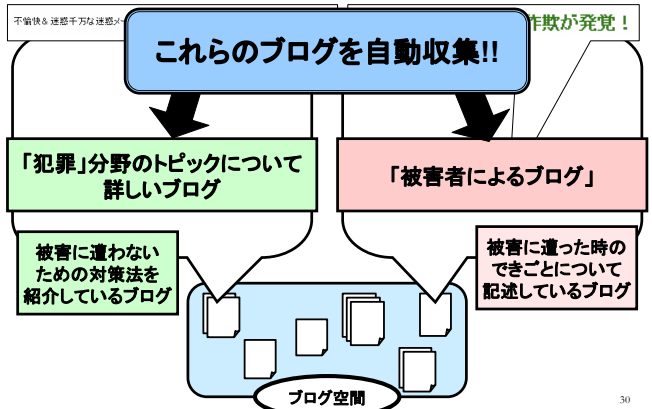
28

「犯罪」分野に関連するブログの例



29

本研究の目的



30

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

31

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!**
 消費者が実際に遭遇した幅広い分野のトラブルに関する相談事例や対策方法を紹介

しかし…

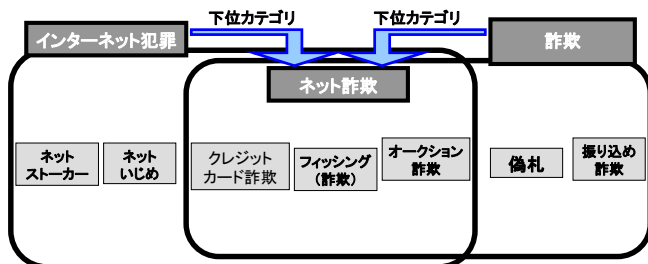
- 公開されている情報は**人手による作業が必要でコストがかかる。**
- **人手による作業のため、公開されている情報の種類・特徴・数が限られている。**

名古屋市消費生活センター

⇒ **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!**

32

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

33

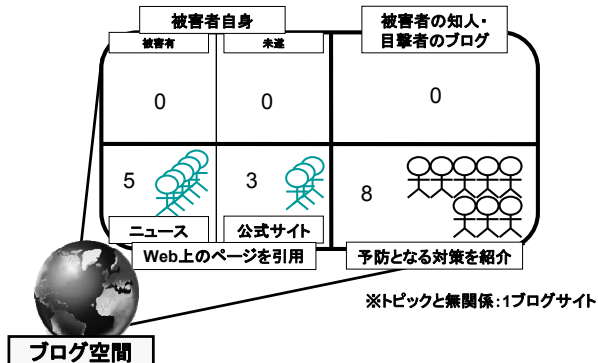
目次

- 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

34

結果:「フィッシング詐欺」

収集したブログサイト数(11ブログサイト)



35

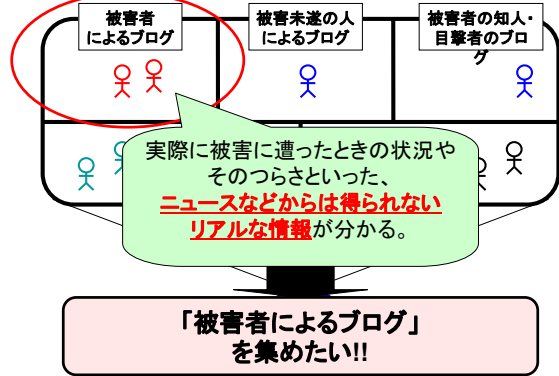
目次

- 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

36

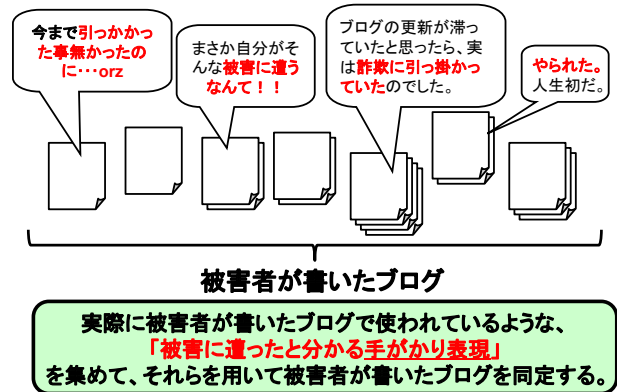
「被害者によるブログ」に注目する理由

「犯罪」分野におけるブログの立場



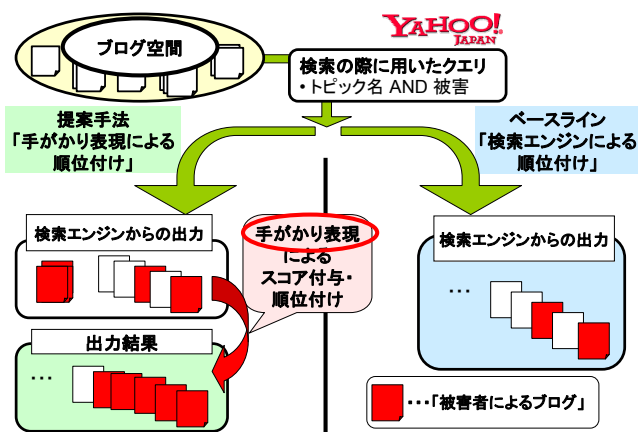
37

「被害者によるブログ」を見つけるために



38

提案手法と検索エンジンとの比較

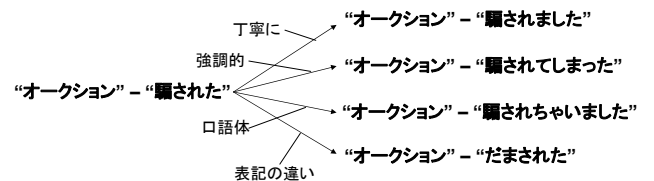


係り受け関係について

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



40

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」ことを表しているような言葉も、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!!」”

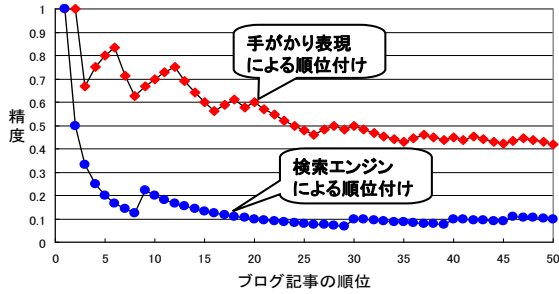
41

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。

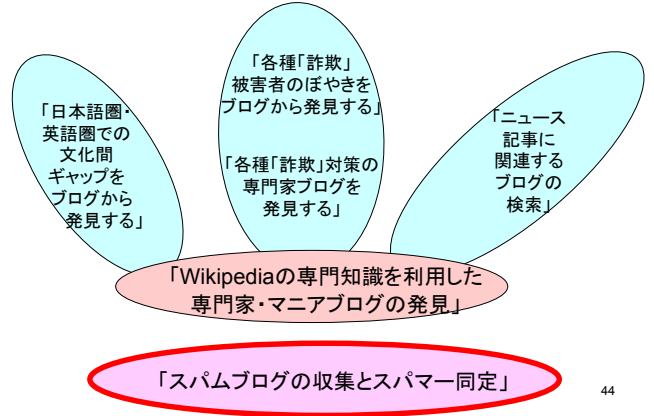
手がかり表現の種類	手がかり表現の例	スコア	種類数			合計
			活用有り	活用無し	その他	
係り受け関係	基本形	10	13	3	3	19
	派生形		84			90
文節単位の表記パターン	高スコア	2	11	1	1	13
	中スコア	1	15	91	7	113
	低スコア	0.5	5	13	0	18

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



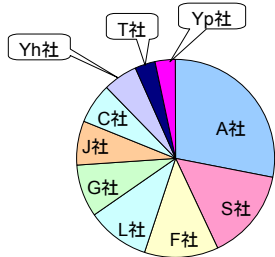
発表の内容

- 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

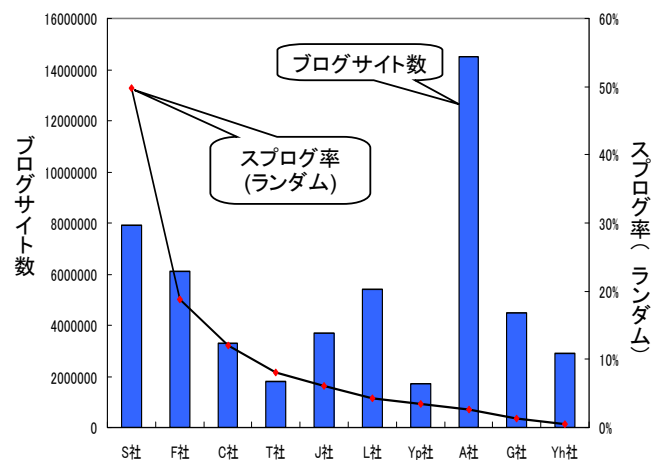
日本語ブログ空間

- 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%)

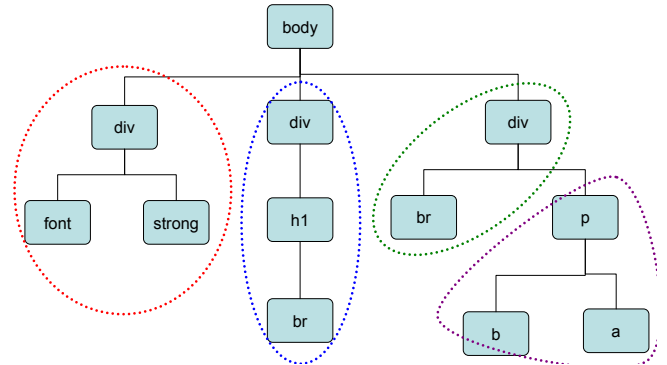


データセット

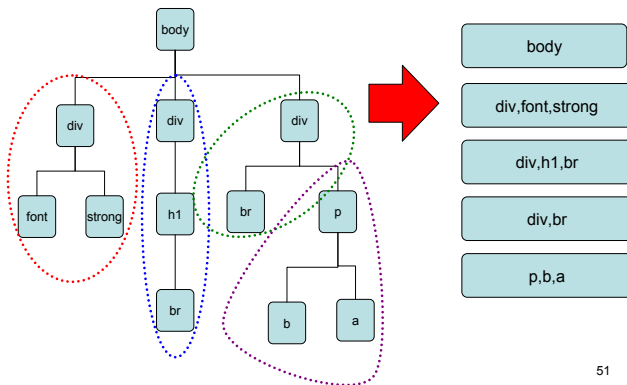
- 520万URLのうち10分の1の52万URLを収集
- ランダムサンプリングにより
各ブログ会社: 500サイト
- それぞれスプログ、非スプログのラベルづけ
- 同じコメントが存在するものは大量生成型スプログ

49

HTMLの木構造をブロックに まとめあげる(ブログサイトs)

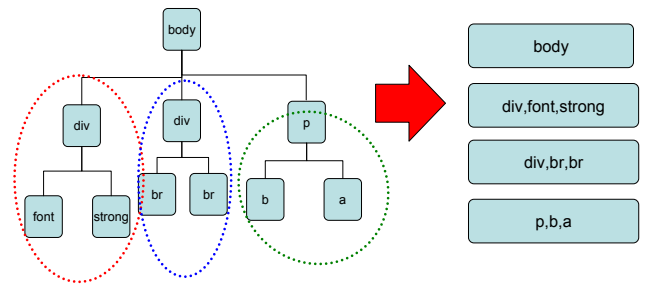


ブロック系列の抽出(ブログサイトs)



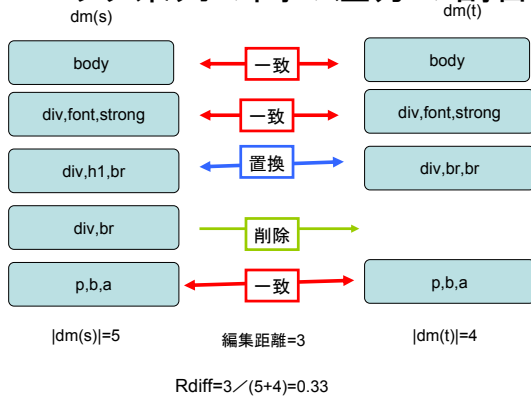
51

ブロック系列の抽出(ブログサイトt)



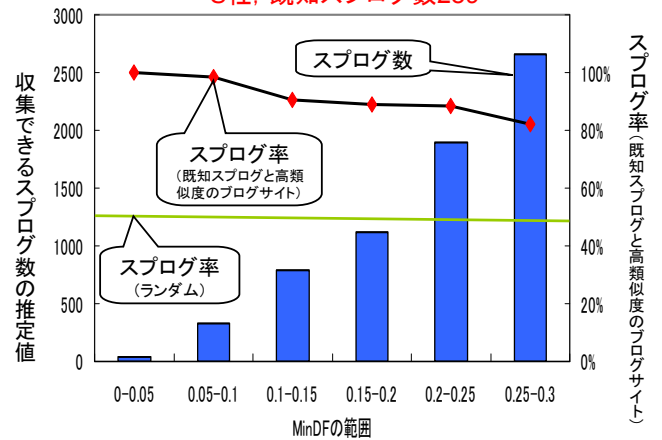
52

ブロック系列の間の差分の割合

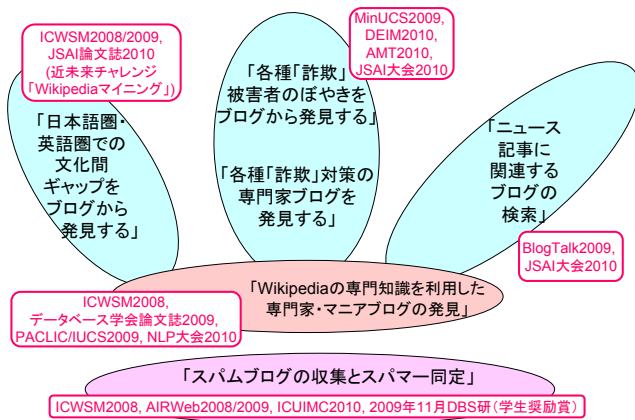


53

S社, 既知スプログ数289



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