

スパムブログに関する定量的調査支援ツールの開発

芳中 隆幸[†] 福原 知宏^{††} 増田 英孝[†] 中川 裕志[‡]

[†] 東京電機大学 未来科学部 ^{††} 東京大学 人工物工学研究センター [‡] 東京大学 情報基盤センター

1 はじめに

今日、ブログツールやブログサービスの普及に伴い、多くの人々がブログサイトを開設し、情報発信できるようになった。一方、ブログサイトの中には価値の低いブログサイト(スパムブログ, Splog(スプログ))が増加し、検索エンジンにおける不当な順位操作や検索結果における精度低下の原因となっている。

Kolari らは英語圏の Splog 空間について調査を行っている [1] が、日本語圏の Splog 空間は英語圏とは異なる傾向にある。また、日々新たな種類の Splog が出現し、いたちごっこの状態が続いている。効果的な Splog フィルタリングの実現には、Splog 空間についての十分な知見が必要である。

本研究では日本語 Splog 空間に関する知見の獲得を目標とし、Splog 空間を定量的に調査するための支援ツール SplogExplorer を開発した。本ツールは 3 つのサブシステムから構成されており、それぞれのサブシステムが Splog 空間を分析するための種々の機能を提供する。

本論文の構成は次の通りである。2 では開発した定量的 Splog 調査支援ツール (SplogExplorer) について述べ、その効果について検証する。3 では本論文の議論をまとめ、今後の課題について述べる。

2 SplogExplorer: 定量的 Splog 調査支援ツール

本研究で開発した定量的調査支援ツール SplogExplorer は以下の 3 つのサブシステムから構成される。また全体の構成を [図 1] に示す。

1. SplogAnalyzer
2. SameArticleIdentifier
3. SplogChecker

以下、各サブシステムの機能と役割について述べる。

2.1 SplogAnalyzer

SplogAnalyzer(SA) システムは利用者がブログ検索を行い、ブログサイトやブログ記事の特徴量を把握す

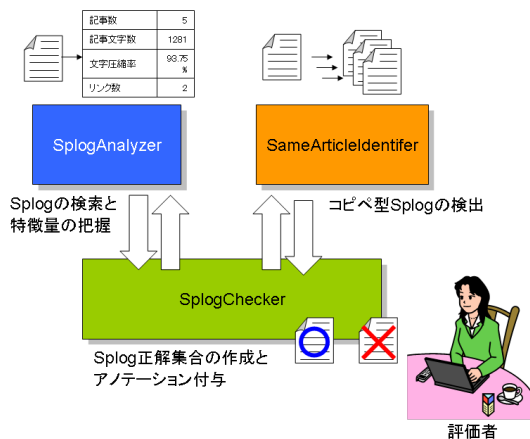


図 1: SplogExplorer システム全体構成

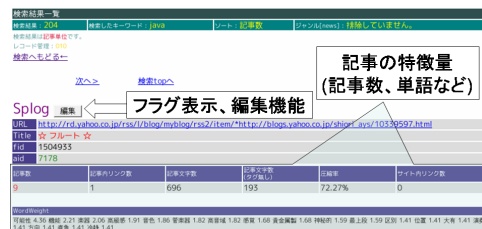


図 2: SplogAnalyzer における検索結果

ためのシステムである。

SA システムには Splog 検索を行うための様々な検索機能が備えられており、キーワードによる検索の他、各ブログサイトの記事数やリンク数などの数値情報を用いた検索を行える。検索結果には実際のブログ記事が持つ記事本文やタイトルが示される他、記事数やリンク数などの特徴量が示される [図 2]。表 1 に SA システムが提示する特徴量を示す。これらの特徴量を提示することで、Splog 空間と Splog でない空間との特徴量の相違を分析することができる [2]。そのほか実際に検索されたブログサイトに対して、SplogChecker(2.3 参照)によって付与された Splog 判定情報も提示され、検索結果上での Splog 空間が持つ特徴量も利用者に提示される。

2.2 SameArticleIdentifier

SameArticleIdentifier(SAI) システムは、他のブログサイトの記事を無断引用するコピー & ペースト型(コ

[†] Takayuki Yoshinaka

^{††} Tomohiro Fukuhara

[†] Hidetaka Masuda

[‡] Hiroshi Nakagawa

School of Science and Technology for Future Life, Tokyo Denki University ([†])
Research into Artifacts Center for Engineering, The University of Tokyo (^{††})
Information Technology Center, The University of Tokyo ([‡])

表 1: SplogAnalyzer にて提示される特徴量

記事数	ブログサイトが一日に書く記事数
記事内リンク数	ブログ記事内のみでのリンク数
全文字数	ブログ記事内の文字数
タグ無し文字数	ブログ記事内のタグを除いた文字数
圧縮率	タグ無し文字数/全文字数
サイト内リンク数	ブログサイト全体での外部リンク数

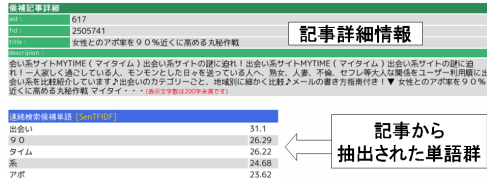


図 3: SameArticleIdentifier における記事解析

パイプ型)Splog を検出するシステムである。ここで SAI システムにおいてコピペ型 Splog を検出するためのアルゴリズムを説明する。このアルゴリズムではまず、コピペ記事の元になる記事を解析する。記事解析時の状態を [図 3] に示す。ブログ記事に対し形態素解析を行った後、抽出された単語に TFIDF 法 [3] を用いた重み付けを行う。重み付けされた上位の単語から順に連続的な AND 検索を行っていき、ブログ記事の絞り込みを行うことでコピペ型 Splog の検知を可能としている。実際にこのアルゴリズムでコピペ型 Splog が検知された結果を [図 4] に示す。また TFIDF 法を使用せず、ある 1 単語と隣合う単語を解析時に抽出することで部分的なコピペにも対応したシステムとなっている。

2.3 SplogChecker

SplogChecker(SC) システムは効率的な Splog の正解集合作成支援を目的としたシステムである。

SplogAnalyzer や SameArticleIdentifier では検索結果からのみ Splog フラグを付加することができるが、本 SC システムでは Splog フラグの付加を一括で行うことができ、効率の良い Splog フラグ編集が行える。また本 SC システムはブログ記事集合全体における Splog の割合と Splog 空間の特徴量も提示する。

[図 5] に SC システムのユーザインタフェースを示す。利用者はアカウントを作成しログイン機能を使用することで Splog フラグ編集回数や記事に対するアノテーション機能を使用できる。また他の利用者による記事へのコメントも閲覧でき、利用者は自分の評価だけでなく他の利用者の評価も参考にしながら Splog 空間の範疇を決定できる。

利用者がどのようなブログ記事やサイトを Splog と判定するかは利用者一人一人によって異なる。そのた

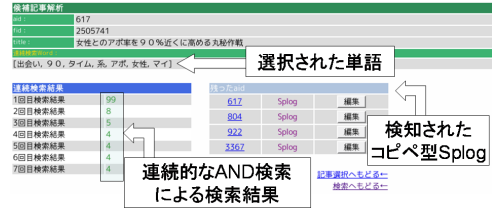


図 4: SameArticleIdentifier による連続的な AND 検索による検索結果

め本システムでは複数の評価者が Splog か Splog ではないかを判定できるような設計となっており利用者側で Splog の定量的定義を決定可能な設計となっている。

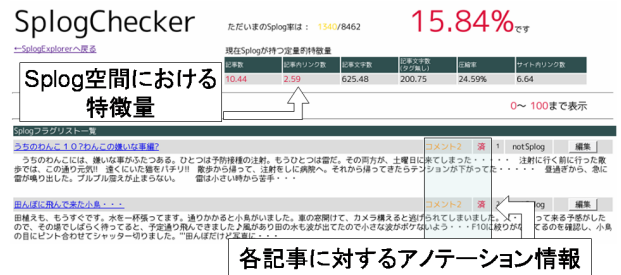


図 5: SplogChecker のユーザインタフェース

3 まとめと今後の課題

本研究では Splog 空間における定量的な調査を支援するためのツール SplogExplorer を開発し、提案ツールが Splog 空間に対しどのような支援をするのかについて述べた。提案ツールは 3 つのサブシステムで構成されており、これらは全て Splog 空間を定量的に調査することを目的としている。

本研究の今後の課題は、データセットの大規模化に向けたツールの構築と精度調査である。また本研究の最終目標は Splog 空間の定量的分析に基づく効果的な Splog フィルタリングの実現であり、今後、本ツールを用いた Splog 空間の分析を行う。

参考文献

- [1] Pranam Kolari et al. Detecting spam blogs: A machine learning approach. *Ph.D. Dissertation*, Dec 2007.
- [2] 石田和成. スパムブログの定量的調査と分離の試み. データベースと Web 情報システムに関するシンポジウム DBWeb2007, No. 5B-3, Nov 2007.
- [3] 徳永健伸. 情報検索と言語処理, 言語と計算, 第 5 巻. 東京大学出版会, 1999.