システム情報工学研究科修士論文概要

年 度	平成 24 年度	学位名	修士(工学)
専 攻	知能機能システム 専攻	著者氏名	森尻 惇宜史

指導教員氏名 宇津呂 武仁

論文題目

HTML 構造の特徴を利用したスパマーの同定

論文概要

個人のブログサイトに含まれる多くの意見情報は，市場の動向を推測する際に有用であるとして近年注目を集めている．一方で，ブログサイトの作成と配信は容易になっており，それによりアフィリエイト収入を得ることのみを目的とするスパムブログ(以下，スプログ)が急増している．スプログにおいては，より多くのアフィリエイト報酬を得るために，一人のスパムブログ作成者(以下，スパマー)が多数のスプログを生成している．それらのスプログは，プログラムを用いて，機械的な文書作成や他サイトの文書を引用することで記事を自動的に生成し，大量のリンクを有するブログを機械的に自動生成する．このようなスプログの存在は，情報検索品質の低下などの問題を引き起こすため，スプログの分析や検出を目的とした研究が進められている．既存の研究としては，言語情報，リンク情報，HTML タグ情報，時間情報といった多様な特性を手がかりとして，スプログを検出する手法が提案されている．

本論文では，ブログサイトの持つ HTML 構造の特徴を利用することで，ブログサイト集合の中から，同一スパマーにより生成された可能性のあるスプログの候補を効率的に収集する手法を提案する．スパマーがプログラムを用いて多数のスプログを生成する場合，生成されたスプログ同士の HTML 構造には共通の断片的特徴が散見されることになる．そこで，そのような断片的特徴に着目してブログサイトを収集することで，同一スパマーにより生成されたスプログを収集することができる．具体的な手順としては，まず，ブログサイトの HTML 文書中に存在するタグを用いて DOM 系列を抽出する．その DOM 系列の部分列を断片的特徴とし，同じ断片的特徴を持つブログサイトを収集する．これにより，HTML 構造に同じ特徴を持つブログサイトを収集することができる．更に，収集したブログサイト集合中で個々のブログサイト同士を組として DOM 系列全体を比較し，似た DOM 系列を持つブログサイトが多く含まれるブログサイト集合のみを抽出する．一般的に，異なる作成者により作成されたブログサイト同士の DOM 系列は似通わないため，これにより，より高い精度で同一スパマーに生成されたスプログのみを収集することができる．以上の手法を用いることにより，高適合率で同一スパマーにより生成されたスプログが同定できることを示す．

審査日 平成 25 年 1 月 30 日

審査員	(大学名 職名)	(学位)	(氏名)
主査	筑波大学 教授	博士(工学)	宇津呂 武仁
副査	筑波大学 教授	工学博士	白川 友紀
副査	筑波大学 准教授	博士(工学)	古賀 弘樹