審 査 の 結 果 の 要 旨

氏名　　鄭　容朱

本論文は「A Study on the Evolution and Emergence of Web Spam （ウェブスパムの進化と出現に関する研究）」と題し、英文8章から構成されている。検索エンジンのランキングを故意に操作しようとするウェブスパムについて、複数年にわたる日本の大規模ウェブアーカイブを用いた分析を行い、ウェブグラフ上におけるリンクスパムの分布、スパムのトピック分布、及びそれらの経時的変化を明らかにすると共に、新たなスパムへのリンクが発生する個所を検出する手法を提案し、その有効性を示している。

第1章は、「Introduction (序章)」であり、本論文の背景および目的について概観し、本論文の構成を述べている。

第2章は、「Spamming Techniques （スパム手法）」と題し、ウェブスパムが攻撃対象とする、検索エンジンのランキング手法と、それらに対する典型的なスパムの手法についてまとめている。

第3章は、「Related Work （関連研究）」と題し、ウェブスパムの対策手法に関する関連研究をまとめている。

第4章は、「Analysis of Link Farm Distribution （リンクファームの分布に関する分析）」と題し、ある時点でのウェブグラフにおけるリンクスパムの分析手法を提案すると同時に、実際のウェブアーカイブを用いた分析結果を示している。リンクスパムが構成する稠密なリンク構造であるリンクファームを、ウェブグラフから抽出する再帰的強連結成分分解手法を提案し、一定以上のホスト数を持つ強連結成分が実際にスパムであることを示すと共に、ウェブグラフにおけるリンクファームの分布を示している。さらに、リンクファームのトピックを、URL情報を用いて精度良く分類する手法を提案し、スパムのトピック分布を示している。

第5章は、「Analysis of Link Farm Evolution （リンクファームの進化に関する分析）」と題し、第4章で抽出したリンクファームの経時的な変化を、複数年にわたるウェブアーカイブを用いて分析している。分析結果から、リンクファームに含まれるホスト数は、多くの場合大きく変化せず、減少する場合が多いことを見出した。これは、リンクファームの多くは、一度作られると放置されるか破棄されることを意味する。また、スパムのトピック分布については、全体的な傾向に変化は見られないが、内容が動的に変化することを明らかにした。

第6章は、「Link Hijacking Detection （リンクハイジャックの検出）」と題し、スパム作成者がブログのコメント欄、掲示板、破棄されたドメインの買い取りなどを通じて通常のホストからの被リンクを得る、リンクハイジャックと呼ばれるスパム行為を検出する手法を提案し、ウェブアーカイブに適用した結果を示している。ハイジャックされたホストは継続的に攻撃されることが多いため、新たに出現するスパムを観測する拠点となり得る。提案する手法は、各ホストに対して、ページランクアルゴリズムに基づく2種類のスコア、即ちホワイトスコアとスパムスコアを算出し、スコアの逆転が多く生じる場所を特定するものであり、実際のウェブアーカイブに本手法を適用し、上位200件において約70%の精度でハイジャックされたホストを検出できることを示している。

第7章は、「Spam Link Generator Identification （スパムリンクの発生個所の特定）」と題し、複数年に

わたるウェブアーカイブを用いて、実際にスパムへのリンクを増加させているホストを特定する手法を提案し、その結果を示している。スパムリンクを継続的に発生させている個所を特定できれば、新たに出現するスパムを観測することが可能となる。本手法では、リンクハイジャック検出で用いた、ホワイトスコア、スパムスコア、及び隣接ホストのURLに出現する文字列、リンク数の経時的変化などを特徴量とした分類器をオンライン学習により構築する。実際に複数年にわたるウェブアーカイブを用いた実験を行い、高い精度でスパムリンクの発生個所を特定できることを明らかにした。また、特定した多くの発生個所が、現在もスパムリンクを生成していることを確認している。

第8章「Conclusions （結論）」では、本論文の成果と今後の課題について総括している。

以上これを要するに、本論文は、ウェブ検索エンジンのランキング操作を目的としたスパム行為に対して、ウェブグラフ上におけるリンクスパムの抽出手法及びトピック分類手法を提案すると共に、リンクスパムの分布、スパムのトピック、及びそれらの経時的変化を分析した結果に基づいて、スパムへの新たなリンクの発生個所を精度良く特定する手法を提案するものであり、複数年にわたる大規模な日本のウェブアーカイブを用いた実験により、リンクスパムの様々な性質を明らかにすると同時に、新規スパムへの有効な対応策に繋がる知見を与えており、情報理工学上貢献するところが少なくない。

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。