

平成 21 年 4 月 21 日現在

研究種目：特定領域研究

研究期間：2006～2010

課題番号：18049013

研究課題名（和文）

情報爆発時代におけるサイバー空間情報定量評価基盤の構築

研究課題名（英文）

Building Quantitative Evaluation Platform for Exploding Cyber Information

研究代表者

喜連川 優

東京大学・生産技術研究所・教授

研究者番号：40161509

研究分野：総合領域

科研費の分科・細目：情報学・メディア情報学、データベース

キーワード：情報システム、コンテンツ・アーカイブ、計算機システム、データストレージ、情報工学、ウェブ情報、リンク解析、性能評価

1. 研究計画の概要

近年情報が爆発的に増加しており、膨大な情報源から真に必要な情報を如何に抽出するかは、極めて深刻な課題であり、本研究では当該課題に挑戦すし、情報源の中でも最も増加率の高いウェブ情報源に対して定量的評価基盤を構築することを目的とする。即ち、情報獲得に関して種々の研究がなされているが、ウェブでは刻々とコンテンツが変化することから、それらの成果を再現性のある形で定量的に示すことは不可能であった。本研究では、各種手法の有効性を定量的かつ再現性を持たせた形で評価するプラットフォームを構築する。具体的には、既に独自に開発した超高速クローラ（ウェブページ収集システム）を改良し、細粒度時系列差分スナップショットを取得し各種テキストインデクス、種々のリンクインデクス等を実装した高度プラットフォームを構築する。

2. 研究の進捗状況

サイバー空間情報定量評価基盤の基本アーキテクチャ設計を行い、その構築を進め、基本アーキテクチャの部分評価を試みると同時に、再現性のある定量的評価手法を実現するために現在のサイバー空間の解析を行い、定量評価基盤への組込み手法について検討を行った。

1)最新のストレージ仮想化手法を用いた拡張可能なシステムアーキテクチャ並びに格納構造の設計を行うと共に、IP-SAN等のストレージ高速通信機構について検討、実験を行った。

2)定量評価結果の解釈支援を目的とし、主としてサイバー空間構造解析を対象に、サイバーコミュニティ抽出、スパム解析、パーティカル検索のた

めのエンティティ抽出等のツール群を構築するべく研究開発を進めている。スパム構造を分析する新たな構造解析手法の開発を行い、全ウェブ空間におけるスパム分布の概要を明らかにした。また、効率の良いウェブ解析のために、開発してきた超高速ウェブクローラを改良し、各種テキストインデクス、特徴的グラフ様態に着目した種々のリンクインデクス、並びに各種ユーティリティを実装した高度プラットフォームを構築を開始した。

3. 現在までの達成度

当初の計画では、三年度を目指し、サイバー空間解析ツール群の充実とサイバー空間情報蓄積を行う予定であったが、蓄積された情報を用いてスパム構造を分析するに至っており、当初の計画以上に進展している。

4. 今後の研究の推進方策

本研究は順調に推移しており、従来の計画にあるようにサイバー空間情報の蓄積を行うとともに、開発したサイバー空間解析ツールを利用することで、サイバー空間の時間的遷移等新たな軸からの解析を試みたい。

5. 代表的な研究成果

〔雑誌論文〕(計 10 件)

豊田 正史 検索エンジンスパムとその対策技術
人工知能学会誌 23 6 760-766 2008 年

相良 毅,喜連川 優 住所情報を用いた店舗名称の
クリーニング手法 電子情報通信学会論文誌 D
J91-D 3 531-537 2008 年

大塚 真吾,喜連川 優 Web アクセスログとその利
活用 人工知能学会誌 Vol.21 No.4 410-415 2006
年

〔学会発表〕(計 69 件)

Masashi Toyoda, Masaru Kitsuregawa What's
Really New on the Web? Identifying New Pages
from a Series of Unstable Web Snapshots,
WWW2006