様式 C-19

科学研究費補助金研究成果報告書

平成 22 年　6 月　3 日現在

研究種目：　若手研究（Ｂ）
研究期間：　平成 20 年度　～　平成 21 年度
課題番号：　20760255
研究課題名（和文）　ネットワーク情報源符号化に基づく情報検索の性能限界の究明とその応用
研究課題名（英文）　On theoretical bounds of information retrieval based on network source coding and its applications

研究代表者
木村　昭悟　(KIMURA AKISATO)
日本電信電話株式会社 ＮＴＴコミュニケーション科学基礎研究所
研究者番号：　10396202

研究成果の概要（和文）：　本研究では、情報検索をネットワーク情報源符号化の観点からモデル化・定式化し、そのモデル・定式化に基づいて情報検索の性能限界をある情報源のクラスについて理論的に証明すると共に、その性能限界に漸近する性能を達成できる具体的な実現方法を示した。また、上記性能限界を理論的に証明できていないより広いクラスの情報源に対して、性能限界を近似的に導出できる数値計算のためのアルゴリズムを開発した。

研究成果の概要（英文）：This project focuses on deriving theoretical bounds of performances of information retrieval within a framework of Shannon theory, especially network source coding. We have clarified some of those theoretical bounds for a specific class of information sources as well as a method for asymptotically achieving the theoretical bound. Also, we developed an algorithm of deriving approximate theoretical bounds for a broader class of information sources for which the strict theoretical bound has not been derived yet.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,000,000	300,000	1,300,000
2009 年度	600,000	180,000	780,000
年度			
年度			
年度			
総　計	1,600,000	480,000	2,080,000

研究分野：　通信・ネットワーク工学
科研費の分科・細目：　情報理論
キーワード：　情報理論、情報検索、データベース、パターン認識

１．研究開始当初の背景

大容量ネットワークの普及、音楽・映像配信の一般化などにより、膨大な量のメディア情報を取得し保存する時代となっている。それに伴い、所望のメディア情報を高速かつ正確に探し出す技術の開発が各方面で進められており[木村 2002: 信学論 D-II] [Kimura 2008: IEEE Trans ASLP] [Kashino 2007: ICAPPS2007]、検索性能は日々進化を遂げている。一方で、それらメディア情報検索技術の性能限界について、現実的な仮定の下で検討されている研究は少ない。

メディア情報検索技術の性能限界について言及する数少ない研究の１つとして、検索

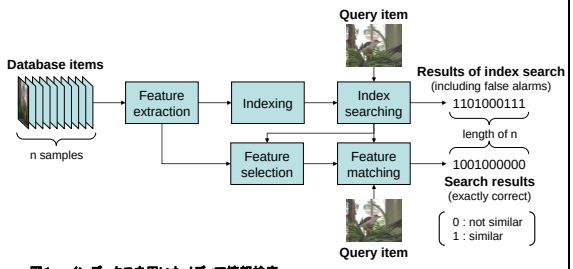

図1：インデックスを用いたメディア情報検索

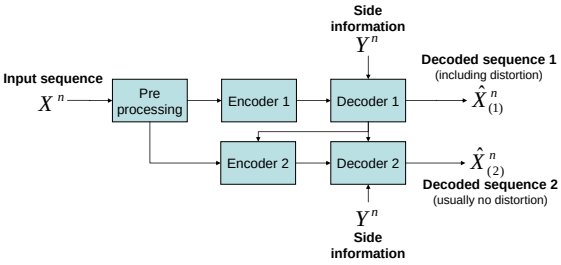

図2：インデックスを用いたメディア情報検索のネットワーク情報源符号化モデル
インデックスを用いた上図のメディア情報検索において、蓄積信号の各サンプルに対するインデックスの付与を、蓄積信号に対応する特徴系列の圧縮として捉えると、下図に示す、歪みを許すネットワーク情報源符号化としてモデル化できる。

高速化の核としてしばしば用いられる多次元インデックス手法 [Beckman 1990: ACM SIGMOD 90'] について、インデックスサイズと検索速度との関係について、理論的解析・実信号による評価が行われている [Boehm 2000: ACM Trans DS] [Tao 2004: IEEE Trans KDE]。しかし、これらの解析や評価は、データベースに登録されている各データの分布が一様であることを本質的な仮定としている、という致命的な欠陥を抱えているため、実世界で妥当性を持つ仮定の下での評価が原理的に不可能であった。

一方で近年、情報理論的なアプローチを用いて、既存研究における本質的な欠陥を原理的に解決する研究が始まりつつある [Tuncel 2004: IEEE Trans IT]。このアプローチでは、データベースに登録されている各データが所定の確率分布に従って生成されることを仮定し、情報源符号化もしくは通信路符号化の枠組で所定の情報検索モデルを定式化して理論限界を導出する。

上記の背景に基づき、研究代表者らは、これまでに、メディア情報検索技術において最も基本的かつ重要な技術の 1 つである、インデックスを用いた情報検索（インデックス検索）を題材として取り上げ、これをネットワーク情報源符号化の枠組でモデル化できることを示した [Kimura 2006: STW2006] [Kimura 2006: SITA2006]（図 1、図 2）。これにより、インデックスをどの程度用意すれば所定の検索精度を担保したまま検索速度をどの程度まで向上できるか等、インデックス検索の性能を測定する問題を、情報源符号化モデルにおいて所定のひずみ上限値を担保したままその程度情報を圧縮できるか、といった、ひずみを許容する情報源符号化における基本問題に置き換えて考えることが可

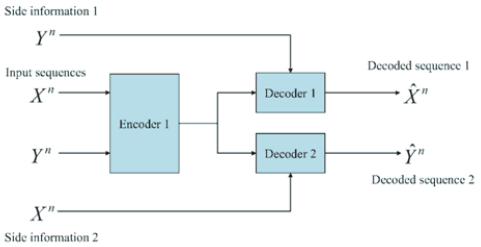

図3：インデックスを用いたメディア情報検索のネットワーク情報源符号化モデルの部分モデル

能となった。

その上で、図 2 の情報源符号化モデルにおける符号化率の理論的な上界と下界を明らかにすることで、インデックス検索の性能について、その理論的な上界と下界を明らかにした [Kimura 2006: SITA2006]。また、図 2 のモデルの部分モデル（図 3）について、ネットワーク情報源符号化としての理論的な限界を完全に明らかにすると共に、その理論限界に漸近する符号の具体的な構成方法を示した [Kimura 2007: IEICE Trans Fundamentals]。これにより、理論限界に迫る性能を保持するインデックス検索のアルゴリズムを構築する上での指針の 1 つを与えた。

２．研究の目的

これまでの研究経緯と議論を踏まえ、本研究では、以下の 4 点について、理論と実証の両面から詳細に検討を進める。

(1) これまでに導出したインデックス検索性能の理論限界を、具体的に与えられたデータ分布について解析的もしくは数値計算によって算出する。具体的には、検索漏れを生じないという前提の下で、検索速度とそれを実現するために用意するべきインデックスサイズとの間にあるトレードオフの関係を、定量的に明らかにする。

(2) 第 1 項目で得られた結果に基づき、既存のインデックス手法の性能評価を行う。すなわち、既存のインデックス手法が性能限界にどの程度迫っているかを定量的に明らかにし、どのインデックス手法が優れているかを評価する。

(3) 第 1・2 項目における検討で得られた知見に基づき、既存のインデックス手法と比べてより性能限界に近い性能を実現できる可能性のある、具体的なインデックス検索手法を構築する。

第 1・2 項目の解決により、インデックス検索技術の目指すべき到達点が与えられると共に、以下第 3 の課題の解決につながる重要な知見を得る。

また、続く第 3 項目の解決により、理論的性能限界という共通の評価軸の下で、従来の

インデックス手法によってもたらされる検索性能を本質的に改善することを可能にする。

以上の通り、本研究により、インデックス検索などに代表されるメディア情報検索技術の目指すべき到達点が明らかになると共に、情報検索アルゴリズムを評価する指針が与えられることになり、工学的視点から非常に重要な成果になることが期待される。

３．研究の方法

これまでの議論を踏まえ、本研究では、以下に挙げる 3 つの方向性で検討を進めた。

(1) より一般のネットワーク情報源符号化モデルに対して、その理論的な限界を明らかにすることで、メディア情報検索の性能限界を理論的に解明する足がかりを作る。

(2) 理論的な性能限界が解明されていないより広いクラスの情報源に対して、性能限界を近似的に導出可能にする数値計算アルゴリズムを開発する。

(3) これらの理論的な性能限界を漸近的に達成可能な情報検索アルゴリズムを、ネットワーク情報源符号化の観点から構築する。

４．研究成果

平成 20 年度は主に理論的解析及び実証的側面に必要な理論基盤の整備に、平成 21 年度は主に実証的側面に、それぞれ焦点を当てて研究を遂行した。得られた研究成果とその概要を、以下に記載する。

(1) 情報検索の理論限界の究明及びその限界に漸近する具体的な符号化方法の提案：

これまでに得ていたインデックス検索のネットワーク情報源符号化モデルの部分モデルについて、その部分モデルを一般化した図 4 に示す符号化モデルを形成し、そのモデルについてネットワーク情報源符号化問題としての理論限界を完全に解明した（雑誌論文①）また、これらの理論限界を漸近的に達成可能とする具体的な方法を提案し（雑誌論文①)、一般には理論的限界を達成するためには膨大な計算量と記憶容量が必要であることを示した。

上記の成果及び先行研究 [Kimura 2007: IEICE Trans Fundamen-tals] により、図 3 のモデルが現実的な計算量で理論的限界を達成する方法が存在する特殊例となっていることが示された。しかし、先行研究の方法では、記憶容量に関する問題が解決されずに残っていた。そこで、本研究では、記憶容量をほとんど必要とせずに理論的限界に漸近する性能を実現可能な方法を新たに開発した（学会発表⑤)。

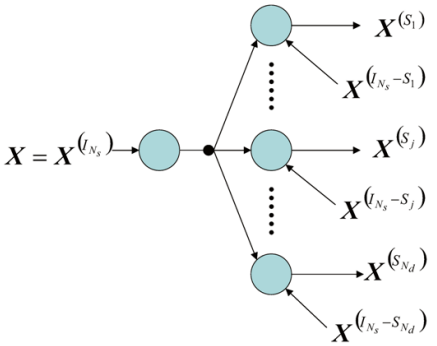

図4： 図3のモデルを一般化したネットワーク情報源符号化モデル

これらの研究成果により、ネットワーク情報源符号化の観点においては、少ない計算量かつごく少ない計算量で理論的限界に漸近するインデックス検索のアルゴリズムを構成できることが明らかになったと共に、その具体的な方法が示された。

(2) インデックス検索の理論的性能限界の数値計算法の導出：

研究代表者らは、これまでの研究で、ネットワーク情報源符号化の枠組を導入することで、インデックス検索の理論的な性能限界に関する上界と下界を示していた [Kimura 2006: SITA2006]。しかし、この理論限界及び前記の成果(1)はいずれも、データが生起する確率分布について、ある特定種類のモデルを仮定していた。それらの確率モデルで記述されるデータ分布については、理論限界を解析的に算出することができるものの、その確率モデルで記述できない、その他多くのデータ分布については、未だに理論限界が知られていなかった。すなわち、これらの理論限界の算出は、どこかの段階で数値計算に頼らざるを得ない。

そこで、本研究では、より広いクラスの確率モデルに対して適用可能な理論限界の算出方法を開発した（学会発表④)。この手法は、制御システムやパターン認識などで事後確率の逐次推定のために用いられている粒子フィルタ（例えば［樋口 2005: 信学会誌］など）の考え方を導入することで、多数のサンプルを用いたシミュレーションで理論限界を探索する数値計算法である。

この成果により、より広いクラスの確率モデルに対してインデックス検索アルゴリズムの性能を評価するための指標を得ることになった。

(3) 情報検索の理論限界に漸近する符号化方法の汎用的な作成方法の提案：

インデックス検索のモデルを含む広いクラスの多端子情報源符号化問題について、情報源に関する統計的性質が未知であっても、

それが既知であるという条件の下で設計された具体的な符号化方法から、理論限界に漸近する符号化法方を作成できる手法を開発した（学会発表①③）。これにより、インデックス検索アルゴリズムとして、これまでに用いられてきたあらゆるタイプの情報源符号を用いることが可能になることが示された。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

① Akisato Kimura, Tomohiko Uyematsu, Shigeaki Kuzuoka, Shun Watanabe “Universal coding for correlated sources over generalized complementary delivery networks,” IEEE Transactions on Information Theory（査読有）, Vol. 55, No. 3, pp. 1360-1373, 2009 年 3 月 1 日.

〔学会発表〕（計５件）

① Shigeaki Kuzuoka, Akisato Kimura, Tomohiko Uyematsu “Universal source coding for multiple decoders with side information,” 採録決定, IEEE International Symposium on Information Theory （査読有）, 2010 年 6 月, Austin, Texas, United States.
② 葛岡成晃, 木村昭悟, 渡辺峻 “Perspectives on multi-terminal source coding; Distributed encoding, distributed decoding, and their applications,” 情報理論とその応用シンポジウム（招待講演）, 2009 年 12 月 2 日, 山口県山口市
③ 葛岡成晃, 木村昭悟, 植松友彦, “Universal source coding for multiple decoders with side information,”, シャノン理論ワークシップ, 2009 年 9 月 24 日, 愛媛県松山市
④ 木村昭悟 “Particle-based simulation of the Gel’fand-Pinsker channel capacity and the Wyner-Ziv rate-distortion function,” 情報理論とその応用シンポジウム, 2008 年 10 月 9 日, 栃木県日光市
⑤ Shigeaki Kuzuoka, Akisato Kimura, Tomohiko Uyematsu “Universal coding for lossy complementary delivery problem,” IEEE International Symposium on Information Theory（査読有）, 2008 年 7 月 11 日, Toronto, Canada.

〔図書〕（計０件）

〔産業財産権〕

○出願状況（計０件）
○取得状況（計０件）

〔その他〕
ホームページ等
http://www.brl.ntt.co.jp/people/akisato/networksource2-j.html
http://www.brl.ntt.co.jp/people/akisato/searchanalysis-j.html
http://www.wakayama-u.ac.jp/~kuzuoka/publications-j.html

６．研究組織
(1)研究代表者
木村　昭悟（KIMURA AKISATO）
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所・メディア情報研究部・研究主任
研究者番号： 10396202