高度分散情報資源活用のための ユービキタス情報システムに関する研究

A Study on Ubiquitous Information Systems for Utilization of Highly Distributed Information Resources

（研究プロジェクト番号：JSPS-RFTF 96P00602）

プロジェクトリーダー
安達　淳　国立情報学研究所情報学資源研究センター・教授

1．研究目的

本研究プロジェクトは、広域ネットワーク上に分散された大規模マルチメディアデータをその所在を意識することなく活用するユービキタスシステムの実現を目的とし、情報の生成から消費に至る各プロセスにあわせて、以下の観点から研究を推進した。

A) 収集フェーズ：各種メディア情報のデータベース構築支援技術
B) 蓄積フェーズ：大規模データの管理技術
C) 提供フェーズ：情報検索技術とユーザインタフェース
D) 活用フェーズ：情報分析とコーパス作成

2．研究成果の概要

2.1　情報収集フェーズにおけるデータベース構築支援

ユービキタス情報システムを構築する上で、既存のメディアを用いて流通している情報を広域ネットワーク上の流通に適した形式に変換することは非常に重要な課題である。本研究では、紙に印刷された文書、ニュース映像、音楽という３つの情報を例にそれぞれの情報をデータベース化する技術について研究を進めた。

印刷物からのデータベース構築支援に関しては、OCR や文書画像解析技術が開発されているが、本研究では、これらの技術を用いて電子化された情報を効果的に利用する技術に焦点をあてて研究を進め、確率的な文字列マッチング技術に基づいた情報検索法およびその処理を高速化するためのインデキシング法を開発した。本研究で提案した確率的近似マッチング法は、OCR の誤りパターンを表す統計モデルを訓練データから構築することによって、高い検索性能を実現した[7]。

ニュース映像に対するデータベース構築支援では、ニュース映像に含まれた映像情報、映像に付随するテキスト情報を駆使し、映像検索に必要な情報を抽出する手法を開発した。本研究では、特にニュース映像中に現れる人物に焦点を当て、映像中の人物の顔からその人物名を同定したり、逆に人物名からその人物の登場する映像の関連部分を抽出する手法を開発した。

音楽情報に対するデータベース構築支援では、ハミングによる音楽検索を実現するために、隠れマルコフモデルに基づく音楽の構造解析法および問い合わせに用いられるハミングと音楽のフレーズを近似マッチングするためのインデクス作成法を提案した。

これらの技術は、広域ネットワーク上に存在する多様な形態の情報を効果的に検索するためのデータ解析技術であり、ユービキタス情報システムの大規模コンテンツを構築するための基本技術として重要である。

図１　データベース構築支援の概要

2.2　情報蓄積フェーズにおける大規模データの管理

定形大規模テキストデータの管理技術は、これまでに多くの研究が行われ、データベース管理システムに反映されてきた。ユービキタス情報システムでは、不定形マルチメディア情報の管理が求められるため、本研究では、画像等のマルチメディア情報の高速検索に必要な多次元データのインデキシング法を開発した[4]。この技術は、それまでの技術で実現された性能を大きく上回るものであり、その後のこの分野の研究に大きな影響を与えた。

2.3　情報提供フェーズにおける情報検索とユーザインタフェース

情報提供フェーズでのシステムの役割は、多様なユービキタス情報源から、利用者が必要な

情報を効率良く探しだすことを支援する点にあり、本研究では、情報検索精度の向上と多言語検察という２つの観点から研究を進めた。

情報検索精度の向上に関しては、テキストを中心に文書の構造、センテンスの構造、文書集合のクラスタ構造、文書中に存在する各種の構造を用いた高精度検索手法を提案した。

検索精度の向上については、まず、文書全体の大域的構造と文というローカルなコンテキストを用いて語彙の多義性を扱うことを試みた。大域的な情報は、機能構造(背景、目的、結論等)とし抽出され、語の出現位置に応じてその語の役割を考慮する。一方ローカルな情報として文書や問い合わせに含まれるテキストから係り受け構造を抽出し、語の多義性の解消を試みた。後者の手法は、17%程度の検索性能向上を実現した。さらに、文書クラスタを用いて各文書の特徴ベクトルを修正する方法を提案した[6)]。この手法は、語と文書の共起情報から、重要語に基づいたクラスタを作成し、このクラスタ構造を用いて各文書の特徴を修正することによって検索精度を向上させるもので、7%〜11%の検索性能向上を実現した。

多言語文書に対する検索を実現するために、本研究では、語彙レベルでの言語間の対応をコーパスから構築する方法を提案した[5)]。この手法は、専門用語をノード、対訳対をリンクと見て大規模な「対訳」グラフを生成し、グラフをクラスタリングすることによって精度の高い訳語クラスタを作成することに成功した。図２は、この成果を多言語検索に応用することによって得られる検索精度の向上を示している。

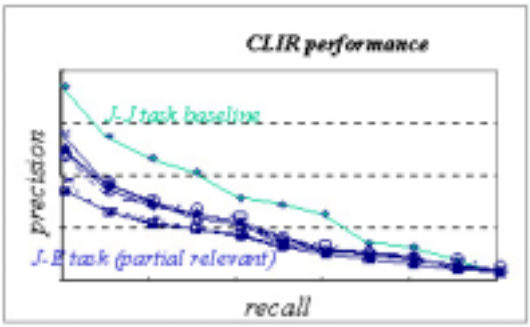

図２　訳語クラスタの多言語検索への応用

2.4　情報活用フェーズにおける情報分析とコーパスの作成

国立情報学研究所が保有する大量なデータを分析し、情報検索手法の評価に必要な問い合わせと正解文書集合からなるテストコレクションを構築した[1,2)]。このテストコレクションは、各種の検索手法を統一的に評価できることが判明したため、本研究プロジェクトに閉じず、情報検索研究コミュニティに公開した。本研究では、研究期間内に２種類のテストコレクションを作成し、このコレクションを用いた検索手法の比較と評価のワークショップを開催し、国内外から多くの研究グループの参加を得た。このテストコレクションは、研究成果の評価に必要なデータベースとして公開されており、情報検索研究コミュニティの研究促進に貢献している。

３．おわりに

本研究では、ユービキタス情報システムを実現するための要素技術について情報収集、蓄積、提供、活用の観点から研究を進めた。本研究で考案された各種要素技術は、マルチメディアデータのインデキシングや情報検索での大幅な性能向上に寄与した。また、それらの技術は、文書画像や映像を対象としたデータベース構築支援システムというユービキタス情報システムのコンポーネントシステムとしてまとめられた。さらに、本プロジェクトの情報分析に関する研究は、情報検索分野における大規模テストコレクションという形で結実し、この分野の研究者に広く公開され、研究推進に役立てられている。

主な発表論文

(1) Kando, N. and Nozue, T. (ed.), “NTCIR Workshop 1: Proc. of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition”, Published by NCSIS, 1999, ISBN: 4-924600-77-6.

(2) Kando, N., Aihara, K., Eguchi, K. and Kato, H. (ed.), “Proc. of Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization”, Published by NII, 2001, ISBN 4-924600-89-X.

(3) Adachi, J., Hara, S., Yasunaga, H., Hasebe, K. and Ishizuka, H., “Academic Library and Contents in Japan”, Literary and Linguistic Computing, Vol.14, No.1, pp.131-145, 1999.

(4) 片山紀生, 佐藤真一, “マルチメディア情報の大規模処理に向けた多次元インデキシング手法の応用”, 電子情報通信学会誌, Vol. J82-D-II, No. 10, pp.385-395, 1999.

(5) 相澤彰子, 影浦峡, “学術文献の和英著者キーワードを用いた類義語クラスタの自動生成”, 情報処理学会論文誌, Vol.41, No.4, pp.1180-1191, 2000.

(6) Kanazawa, T., Takasu, A. and Adachi, J., “A Relevance-Based Superimposition Model for Effective Information Retrieval”, IEICE Trans. on Information and Systems, Vol. E83-D, No.69, pp.57-64, 2000.

(7) Ohta, M., Takasu, A. and Adachi, J., “Reduction of Expanded Search Terms for Fuzzy English-text Retrieval”, Intl. Journal on Digital Libraries, Vol.3, No.2, pp.140-158, 2000.