研究室公開日：3 月 9 日(木) 11 時～，13 時～，15 時～，(16 時以降は予約制で対応可能)

応用数理ネットワーク研究室　C3-303

梅村恭司 教授（umemura@tut.jp），吉田光男 助教（yoshida@cs.tut.ac.jp）

☆どのようなことを勉強するか？

文字列アルゴリズム／機械学習アルゴリズム／ノンパラメトリック統計処理
さらに，英語／プレゼンテーション／技術文章の書き方

☆テーマの考え方

ユニークな技術／最新の技術を習得して前人未踏のニーズを開拓する。

☆どのようなアプローチで研究するか？

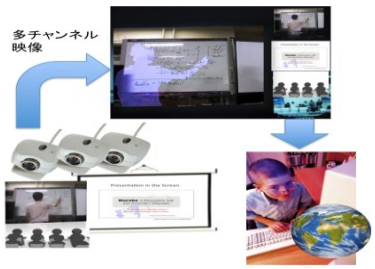
統計処理とアルゴリズムを道具として，蓄積された大量のデータを扱う。
データマイニングで使われているアプローチを踏襲して研究する。

☆キーワード

応用統計学，情報検索，自然言語処理，データマイニング，計算社会科学，e ラーニング

○映像コンテンツのネットワーク応用システム

映像をリアルタイムで処理し，ネットワークで転送し利用するという枠組みのなかで，教育システムへの応用を追求している。要素技術として，前景／後景のリアルタイム分離，映像のスケーラブルな配信技術，マルチチャンネル映像の同時制御などが実現している。この枠組みのうえに，コンピュータを利用したリッチなコンテンツを扱うコミュニケーションについて，時代の先取りをすることも念頭におきながら，教育に応用することを軸足にして研究を進めている。



ネットワーク時代の教育のサポート

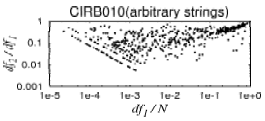
○統計的な手法に根ざした高度なアルゴリズム

ある単語が重要であるかの判断は，コンピュータにとって難しいことのように思えるが，「重要な単語は，ある部分に繰り返して出現する」という性質を使うことで統計的な手法で実現できる。しかし，これを実装するに，分析の対象となる文字の数が多い。高速なアルゴリズムを実装することにより，すべての部分文字列のすべての統計値が，実質的に文字列の長さに比例する程度の時間で表にできるアルゴリズムを実装した。これが，競争力のあるソフトウェアの開発につながっており，共同研究先で製品化されている。

計算量：長さNのすべての部分文字列はN(N-1)/2 個
それぞれについて、文書頻度を求めたい。

実現した方法　前処理：O(N log(N))，文字列ごとに：O(log(N))

x	df(x)	df2(x)
ロ	124696	79894
ロボ	3672	2413
ロボッ	3320	2237
ロボット	3320	2237
ロボットに	577	96
ロボットにつ	30	1
ロボットについ	30	1
ロボットについて	30	1



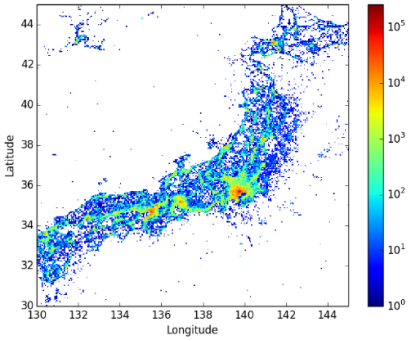
df2(x)：文字列xを2回以上含む文書数

図2　全文字列を対象とした分析

高速な文字列統計値の計数法

○ウェブ／ソーシャルメディアの大規模データ活用

ソーシャルメディア（Twitter）が日常的に利用されるようになり，そこには様々な情報が投稿されている。世界中の位置情報付きツイート（35TB）とリツイート（290TB）をほぼ全て収集するシステムを構築し，これらのデータの活用を検討している。例えば，ツイート内容やフォロー関係（ソーシャルグラフ）から投稿位置や居住地を推定したり，絵文字の利用方法の文化的背景を分析したりしている。さらに，ウェブデータと組み合わせ，注目されている学術情報を分析するシステム（altmetrics.ceek.jp）や，科学報道記事とその根拠となる学術情報とを関連付けるシステムの開発に取り組んでいる。



ツイートの地理的分布