2004－AVM－44　(4)
2004／3／4

モバイル環境におけるユーザクラスタリングを用いた情報推薦システムの検討

金田　瑞規† 　渡辺　裕†

† 早稲田大学 大学院 国際情報通信研究科
〒 169–0051 東京都新宿区西早稲田 1-3-10

E-mail: †mizuki@tom.comm.waseda.ac.jp

あらまし　近年のモバイル端末の普及や，GPS 等の位置測位デバイス搭載機種の登場や処理機能の高性能化からモバイル端末に対する情報推薦システムが注目を集めている．しかし，モバイル端末は表示能力に限界があるため，効率的な情報推薦が不可欠となる．ユーザに対して効率的に情報を推薦する技術として，ユーザの評価履歴を用いて自動的にコンテンツを推薦する「協調フィルタリング方式」が現在最も一般的である．しかし，この協調フィルタリング方式には処理時間等いくつかの問題点が報告されている．特に，モバイル環境におけるシステムでは処理時間が非常に重要な問題となる為，処理時間の性能向上手法が必要になる．そこで本稿では，モバイル環境において効率的な情報推薦を行うシステム構築を目的として，ユーザクラスタリングを用いた協調フィルタリング方式を高速化する手法を提案した．シミュレーションを用いて処理速度，予測精度の検証を行った結果，提案方式の有効性が確認できた．

キーワード　情報推薦システム，モバイルアプリケーション，協調フィルタリング，クラスタリング

A Study on Recommender System using User Profile Clustering for Mobile Application

Mizuki KANADA† and Hiroshi WATANABE†

† Graduate School of Global Information and Telecommunication Studies,Waseda University
Nishiwaseda 1-3-10, Shinjuku-ku,Tokyo, 169–0051 Japan

E-mail: †mizuki@tom.comm.waseda.ac.jp

Abstract　A recommender system for mobile terminal has been paid attention because of wide spread of mobile terminal. Current cellular phones, PDA, and new models equipped with GPS provide high processing capability. However, some suitable recommendation method is needed because of limitation of display size. Collaborative filtering using user profile is the most successful technology for personalized information filtering as an effective recommendation method. However, several problems, such as processing speed, are reported for this approach. In mobile application, this is an important issue. Thus, we propose a method to speed up collaborative filtering using user profile clustering to establish recommender system for mobile application. Through the simulation, we can confirm the validity of our proposed method in processing speed and prediction accuracy.

Key words　Recommender system, Mobile application, Collaborative filtering, Clustering

1. はじめに

近年のモバイル端末の普及や，GPS 等の位置測位デバイスの搭載機種の登場，処理機能の高性能化からモバイル端末に対する情報推薦システムが注目を集めている．昨今の調査では，次世代携帯電話で魅力を感じるサービスの筆頭として，ロケーションサービス等の情報推薦システムがあげられている [1] [2]．一般的に，モバイル端末はその携帯性という特徴から，表示能力には限界がある．したがって，モバイル環境において情報推薦システムを構築するためには効率的な情報推薦が不可欠となる．

効率的な情報推薦技術として，各ユーザに適した情報を提示する「パーソナライゼーション技術」というものがある．これには，静的マッチング方式やルールベース方式等様々な手法があるが，現在最も普及しているものに,「協調フィルタリング方式」という手法が存在する．この協調フィルタリング方式は，ユーザの利用履歴から求める相関を利用して情報推薦を行う手法である．しかしながら，協調フィルタリング方式には予測精度や処理時間といった様々な問題点が報告されている [3]．特にモバイル環境においては処理時間は非常に重要な問題となる為，予測精度処理速度を向上させる手法が必要になる．

そこで本稿では，予測精度を落とさずに協調フィルタリング方式を高速化させる手法として，ユーザクラスタリングを利用した手法についての検討と方式提案およびシミュレーションによる評価結果を報告する．

2. 協調フィルタリング方式

2.1 協調フィルタリング方式とは

パーソナライゼーション技術として，協調フィルタリング方式の他に **1)** 静的マッチング方式や **2)** ルールベース方式というものがある．これらはユーザ自身に嗜好等の属性情報を登録してもらったり，推薦対象へ手動で属性付加を行い，その情報に基づいて情報を推薦する手法である．しかし，これらの手法には推薦対象への属性付加に対するコストや，ユーザ情報入力をユーザ自身に強制するコストやプライバシーといった問題点が存在する．

協調フィルタリング方式は，ユーザの利用履歴から自動的にコンテンツを推薦する手法であるために，これらの問題点は存在しない．そのため，協調フィルタリングは現在最も普及しているパーソナライゼーションエンジンである．これは WWW サイトなどのユーザの集団的な挙動がルールになる仕組みを採っている．具体的には，個々のユーザのアクセス履歴や商品に対する評価などのデータを分析することによりそのユーザと似た因子を持つユーザ集団を探し出し，その集団の挙動を基に対象となっているユーザの嗜好を推測したり，情報を提供したりする．

2.2 協調フィルタリング方式のアルゴリズム

協調フィルタリング方式における，ユーザ a のコンテンツ j に対する推薦基準となる値は，式 (1) の様になる [4]．

$$P_{a,j} = \bar{v_a} + \kappa \sum_{i=1}^{n} w(a,i)(v_{i,j} - \bar{v_i}) \tag{1}$$

$$w(a,i) = \frac{\sum_j (v_{a,j} - \bar{v_a})(v_{i,j} - \bar{v_i})}{\sqrt{\sum_j (v_{a,j} - \bar{v_a})^2 \sum_j (v_{i,j} - \bar{v_i})^2}} \tag{2}$$

ここで，$\bar{v_a}$ はユーザ a の全コンテンツに対する評価の平均値，$w(a,i)$ はユーザ a とユーザ i との相関を表す重み値で，式 (2) であらわされる．$v_{i,j}$ はユーザ i のコンテンツ j に対する評価値，κ は正規化するための係数，n はユーザ数を示す．

ユーザ相関の重み値 $w(a,i)$ は，ユーザの行動履歴や評価履歴を用いて計算され，履歴が似ているユーザの評価値がより重視される仕組みになっている．

式 (1) で与えられる推薦値を基にシステムは推薦を行う．

2.3 問題点

協調フィルタリング方式には，以下のような問題点が存在することが報告されている [3]．

- 誰からのアクセス履歴や評価履歴もないコンテンツは推薦できない
- ユーザの評価データの少ないシステム利用開始時期は正確な推薦ができない
- ユーザ数，コンテンツ数が増加するにしたがって処理時間も増加してしまう．

2.4 モバイル環境への適用

現在の協調フィルタリング方式を用いた情報推薦システムに用いているユーザの評価履歴は，コンテンツに対するアクセス数やユーザ自身がコンテンツを評価した値を用いている．モバイル環境に協調フィルタリングを適用するとこれらに加え，ユーザ移動履歴や場所の情報等も重要な情報として利用できると考えられる．これにより，より柔軟な情報推薦が可能になると予測される．

しかしながら，モバイル環境における情報推薦シ

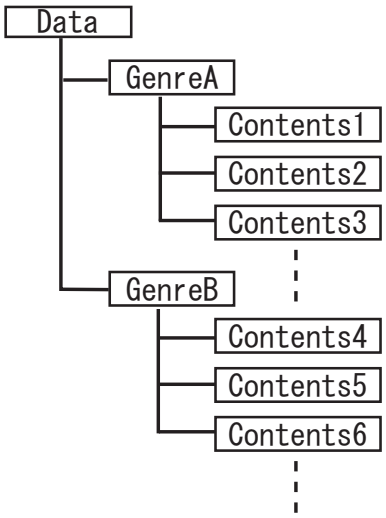

図 1　前提とするデータ構造

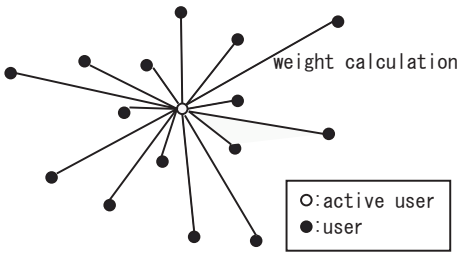

図 2　従 来 手 法

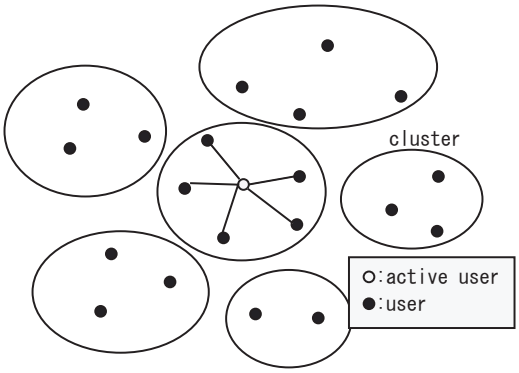

図 3　提案手法 (クラスタリング有り)

ステムでは，システムの処理時間が非常に重要な要素となるために，先に示した問題点の処理時間に関する性能を向上させる手法が不可欠となる．そこで本稿では，予測精度を落とさずに協調フィルタリング方式を高速化する手法として，ユーザクラスタリングを用いた協調フィルタリング方式を提案する．

3. 提 案 方 式

3.1 前 提 条 件

本提案方式で推薦するコンテンツは前提として，コンテンツ自身のデータの上位データとして”ジャンル”等のデータを持ち，図 1 の様な階層的なデータ構造を持つものとする．

現状の情報推薦・検索システムにおいても図 1 のような構造を持っている事が多いため，この前提条件は現実的であると言える．

3.2 クラスタリングを用いた協調フィルタリング

協調フィルタリングにおいて最も処理時間がかかるのは，ユーザ-ユーザ間の相関係数を求める部分である．一方，協調フィルタリングにおいては相関の小さいユーザの評価値は結果にほとんど反映されないという特徴を持つ．したがって，あらかじめ相関の強いと思われるユーザ群をクラスタリングし，相関係数を求めるのはそのユーザ群に対してのみ行えばよく，結果的にシステムにおける処理時間が少なくなると考えられる．

相関の強いユーザをクラスタリングするために用いる特徴量としてはコンテンツの評価履歴が考えられる．しかし，一般的にこのコンテンツ評価履歴には多くの欠損値 (ユーザが未評価) を含んでしまう．また，コンテンツの数だけ次元数があるために，クラスタリングの処理が複雑になってしまう．そのため，ユーザをクラスタリングするのにはこのコンテンツ評価履歴は不向きであると考えられる．そこで本研究では,「コンテンツの評価履歴が似ているユーザは，また上位データ (ジャンル等) に対する評価履歴も似ているに違いない」という仮定のもとで，

（1） 上位データの評価履歴を用いてクラスタリングを行う

（2） そのクラスタ内メンバでのみ協調フィルタリングを行う

という手法を提案する．上位データの評価履歴はコンテンツ評価履歴に比べ欠損値も次元数も少ないためクラスタリングに用いる特徴量として適していると考えられる．

これにより，協調フィルタリングを行うユーザ数が減少する為処理時間が少なくなると考えられる．また，相関の強いユーザをクラスタリングでまとめることにより予測精度の向上も期待できる．図 2，3 に従来手法と提案手法の違いを示す．

3.3 嗜好の変化への対応

本提案手法におけるクラスタリングには，ユーザの評価履歴を用いてクラスタリングを行っている．すなわち，同クラスタ内にいるメンバは似たような嗜好の持ち主であるという事が言える．

したがって，クラスタを固定したままでは，ユーザの嗜好に変化が生じユーザプロファイルが大きく

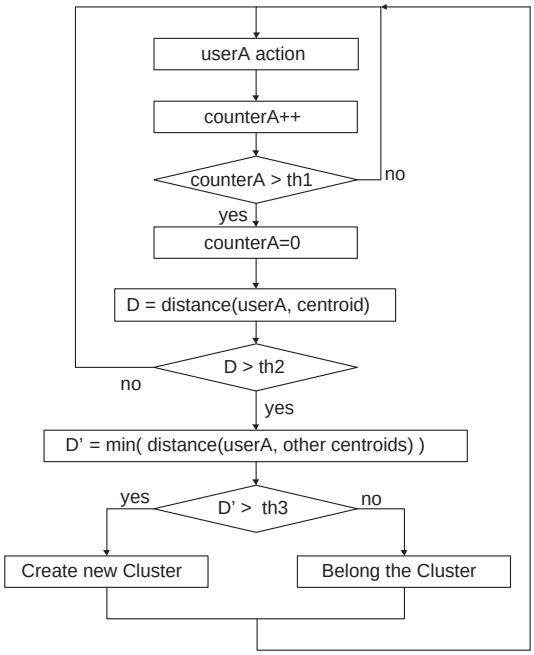

図 4　ユーザのクラスタ参入・離脱の流れ

変化した際において，予測精度が大きく低下してしまうことが予想される．

そこで本研究では，ユーザがセントロイドとの距離に応じて，クラスタからの離脱や参入，新規クラスタの作成を行う手法の提案を行う．本提案手法のアルゴリズムを以下に示す．また，図 4 に提案手法の流れを示す．

（1） ユーザがコンテンツを利用・評価する毎にカウンタを増やす

（2） カウンタの値が閾値 1 以上になったら，所属するクラスタセントロイドとの距離を測定する

（3） クラスタセントロイドとの距離が閾値 2 以下であれば最初に戻り，閾値 2 以上であれば自分の所属するクラスタ以外のクラスタのセントロイドとの距離を測定する

（4） そのうち，距離が最小のクラスタセントロイドとの距離が閾値 3 以下であればそのクラスタに所属し，閾値 3 以上であれば新しいクラスタを作成し，そのクラスタのセントロイドになる．

（5） 1〜4 を繰り返す

4. シミュレーション

4.1 処理時間，予測精度の比較

4.1.1 シミュレーション方法

本提案手法の有効性を検証するために，以下の手順に従ってシミュレーションを行う．システムの評価基準には処理時間及び予測精度を用いる．

本シミュレーションにおいて，クラスタリングアルゴリズムにはクラスタ数を設定するために k-means 法を用いた．

表 1　実験パラメータ

入力モデル	Random 分布，Zipf 分布
ユーザ数	1000，2000，3000，4000，5000
クラスタ数	5，10，20，30
コンテンツ数	100
ジャンル数	5

（1） ユーザ数，入力モデル，クラスタ数を設定する

（2） 1 の条件に従い，ユーザプロファイル (ユーザのコンテンツ評価履歴) を作成する

（3） ユーザプロファイルを正規化する

（4） 全てのユーザが一人一回ずつシステムを利用したと仮定して，通常の協調フィルタリングを行い各ユーザに対する推薦値を求める

（5） 1 で設定したクラスタの数だけ，k-means 法を用いてユーザをクラスタリングする

（6） 各クラスタ毎に協調フィルタリングを行う

（7） 4 と 6 の結果から実行時間，予測精度の 2 点でシステムを評価し従来方式と提案方式の比較を行う．

（8） 1 のパラメータを変更し，シミュレーションを繰り返す

入力データである，ユーザの評価モデル作成には Random 分布，Zipf 分布の 2 種類を用いて作成する．Zipf 分布は WEB のキャッシングシミュレーションに用いられている分布で [6]，嗜好の偏りの強いユーザ群として考えられる．一方の Random 分布は嗜好の偏りの弱いユーザ群として考えられる．

本シミュレーションにおける各パラメータは，表 1 に示す．シミュレーション環境は，CPU：1GHz，メモリ：512MB，OS：WindowsXP である．

4.1.2 予測精度の評価基準

予測精度は，式 (3)〜(6) の 4 つの基準を用いて評価する．ここで，M はコンテンツ数，$p_{i,j}$ はユーザ i に対するコンテンツ j の推薦値，$r_{i,j}$ はユーザ i のコンテンツ j の評価値，$|test|$ はユーザの評価モデルから推薦した結果，$|CF|$ は協調フィルタリングによる推薦結果を表す．

$$MAE = \frac{1}{M}\sum_{j}^{M}|p_{i,j} - r_{i,j}| \tag{3}$$

$$topN = \frac{|test|_{topN} \cap |CF|_{topN}}{N} \times 100 \tag{4}$$

$$precision = \frac{|test| \cap |CF|}{|CF|} \times 100 \tag{5}$$

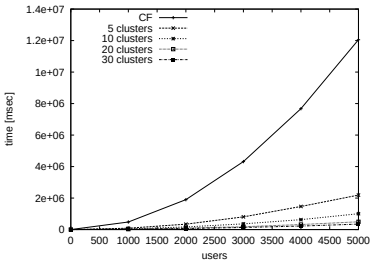

図 5　処理時間 (Zipf)

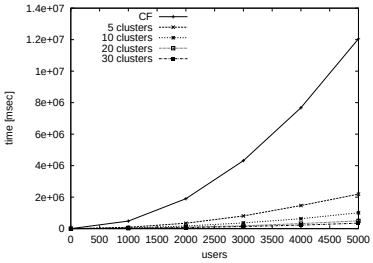

図 6　処理時間 (Random)

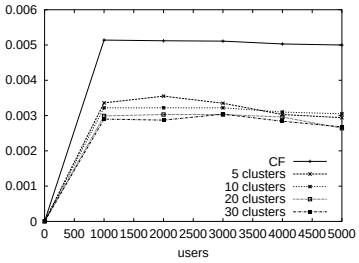

図 7　MAE(Zipf 分布)

$$recall = \frac{|test| \cap |CF|}{|test|} \times 100 \qquad (6)$$

式 (3) の MAE は誤差であり，数値の少ないほうが予測精度が良い．式 (4) の top-N は，予測推薦値の高いほうから N 個のコンテンツを推薦した場合にどれだけ正確な推薦ができたかを表す．式 (5)，(6) の precision，recall はそれぞれ，推薦値が閾値以上のコンテンツを推薦した場合に，システムが推薦したコンテンツはどれだけ正しいか，本来ならば推薦されるコンテンツをシステムはどれだけ推薦できたかを表す．これら 3 つの基準は最高が 100 であり，数値の大きいほうが予測精度が良い．

4.1.3　シミュレーション結果

図 5，6 に処理時間を，図 7〜14 に予測精度を示す．これらの図より本提案方式は Zipf 分布，Random 分布のいづれにおいても従来手法と同程度かそれ以上の予測精度でかつ高速化することが確認された．

4.2　ユーザの嗜好の変化への対応

4.2.1　シミュレーション方法

3.3 章で提案したアルゴリズムを検証するために，

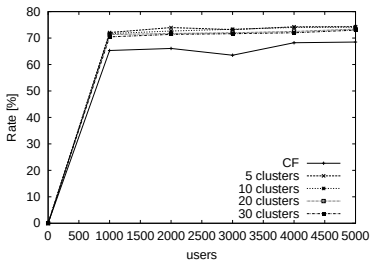

図 8　Top-10(Zipf 分布)

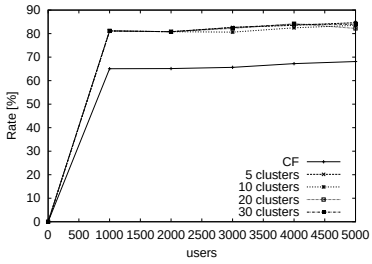

図 9　precision(Zipf)

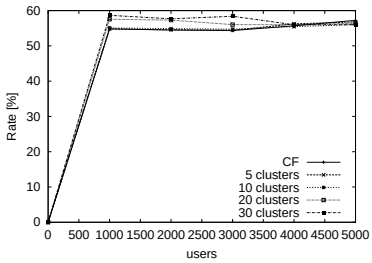

図 10　recall(Zipf)

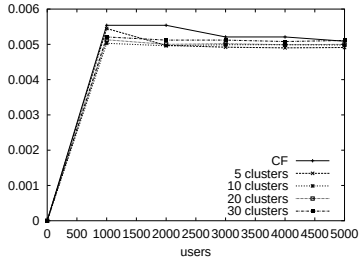

図 11　MAE(Random)

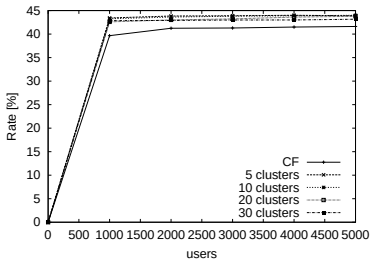

図 12　Top-10(Random)

ユーザ一人の嗜好を意図的に変化させていった場合の予測精度を **1)**:従来方式，**2)**:提案方式 (クラスタ固定)，**3)**:2)+3.3 章のアルゴリズムの 3 手法において測定する．予測精度には MAE を用いる．

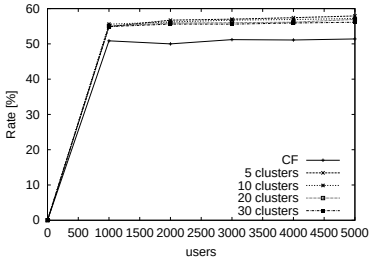

図 13　precision(Random)

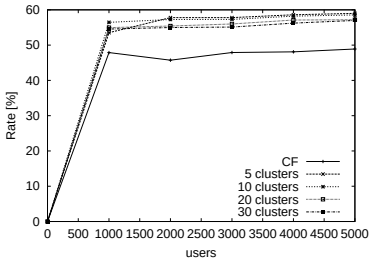

図 14　recall(Random)

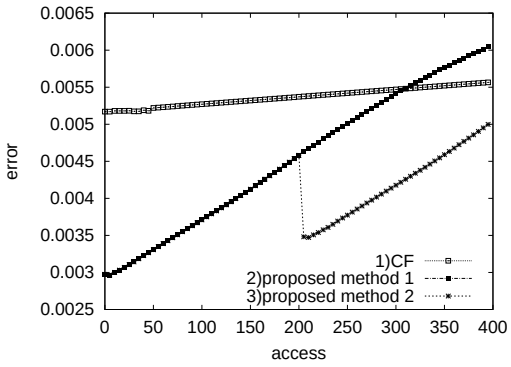

図 15　嗜好の変化による MAE の推移

ここでいう意図的な変化とは，ある特定のジャンルのコンテンツのみにアクセスする事を表す．

本シミュレーションでは，基礎検討のため嗜好を変化させるユーザをひとりのみとし，それ以外のユーザの嗜好には変化はないものとする．

4.2.2　シミュレーション結果

嗜好の変化による予測精度の推移を図 15 に示す．従来方式では MAE はほぼ横ばいとなるが，方式 2) ではクラスタが固定されているために予測精度が回数を重ねる毎に悪化し，従来方式よりも悪くなってしまう．方式 3) ではある程度まで行くとクラスタ移動がおき，再び予測精度が向上していることが確認できる．これにより，本提案方式はユーザの嗜好の変化に対応できる事が検証された．

5.　考　　察

シミュレーションによりクラスタ数を増加させるほど処理時間が短くなることが分かった．しかし，本シミュレーションでは確認できなかったが，クラスタ数を増加させすぎると，クラスタメンバの数が減少し予測精度が落ちてしまう事が予測される．したがって，処理時間・予測精度の面から最適なクラス多数を求める手法を検討する必要がある．

また，ユーザの嗜好の変化への対応は，図 4 の閾値を適当な値を手動で設定して行った．しかし，この閾値の値により予測精度や演算コストなどに変動が出ると思われるので，引き続きシミュレーションを行い，最適な閾値を自動的に求める手法を検討する必要がある．

6.　まとめと今後の課題

本稿では，モバイル環境における効率的な情報推薦システム構築を目的とし，協調フィルタリング方式の問題点を改善する為のユーザクラスタリングを用いた手法を提案した．またシミュレーションによる検証を行い，提案方式の有効性を示した．

以下に今後の課題を示す．

- 全ユーザの嗜好を変化させた場合のシミュレーション
- 実データを用いた評価実験
- 協調フィルタリング方式を用いた情報推薦システムの為の効率的なユーザ及びコンテンツ管理手法の検討

文　　献

[1] 三菱総合研究所, ”次世代携帯電話に関する調査結果, ” http://research.goo.ne.jp/cgi-bin/goo.cgi?::SID=backNumber&_:VP=0101op19/01.html, Mar, 2001

[2] ビデオリサーチ, ”携帯電話の携帯電話の所有と利用状況, ”http://www.videor.co.jp/data/member/marketing/phone/index.htm, 2001

[3] Badrul M.Sarwar 他, ”Using Filtering Agents to Improve Prediction Quality int the GroupLens Research Collaborative Filtering System, ”ACM conference on Computer supported cooperation work, 1998

[4] John S.Breese 他, ”Empirical Analysis of Predictive Algorithm for Collaborative Filtering, ”14th Conference on Uncertainty Artifical Intelligence, 1998

[5] G.Karypis, ”Evaluation of item-based top-n recommendation algorithms, ” Proceedings of the Tenth International Conference on Information and Knowledge Management(CIKM), 2001

[6] M.Aida, T.Nakanishi, ”Design of Address Cache Table for Data Networking Based on Complementary Use of the Two Types of Zipf’s Law, ” Proceedings of the 1997 Asia-Pacific Symposium on Information and Telecommunication Tech., 1997