1P-1

映画推奨システムにおける階層的クラスタリング手法の評価

萩田　盾一†　　辻　秀一‡

東海大学大学院工学研究科†　　東海大学電子情報学科‡

１．はじめに

今日、ADSL や LDSL などの高速なデータ通信技術の登場や加入料の低価格化などによってインターネットに加入する人が大幅に増加し、それにともないホームページを開設する人も増加する傾向にある。そして現在のＷｅｂ情報量は日本だけでも１億ページを超えるまでになっており、インターネットは生活の一部を担う情報ツールとなっている。しかし、インターネットはその膨大な情報量であることから、そこから有益な情報を探し出すといった作業はきわめて困難である。そこで、情報を推奨する技術である「レコメンド技術」とその「レコメンド技術」を実現する為に必要な情報の分類分けを行う技術の「クラスタリング」に注目し研究を行う。

２．従来技術

レコメンドシステム：レコメンドシステムとはユーザーの情報を分析し、各ユーザーの趣向にあった情報を推奨する情報推奨システムであり、例えばインターネット上のショッピングサイトにおいては、顧客の購買データやＷｅｂサイトへのアクセスデータなどから、パタンやルールを検出し、エンドユーザの嗜好や行動を推定し、ユーザのニーズに合致した商品の推薦するシステムである。

クラスタリング：クラスタリングとは複数の特性によって決定された個体間の類似性の指標をもとに、個体の集合をいくつかのグループに分類するための手法である。そしてクラスタリングには階層的クラスタリングと非階層的クラスタリングの大きく分けて２種類の手法がある。さらに階層的クラスタリングには最短距離法、最長距離法等があり、

最短距離法：$D(C_1, C_2)=\max E(x_1、x_2)$

最長距離法：$D(C_1, C_2)=\min E(x_1、x_2)$

というようなアルゴリズムでクラスタの結合が行われる。$D(C_1, C_2)$はクラスタ間の距離、$E(x_1、x_2)$はクラスタ間の類似度である。

Evaluation of hierarchical clustering technique in movie recommendation system

†Junichi Hagita

Graduate School of Engineering,Tokai University

‡Hidekazu Tsuji

School of Information Technology and Electronics,Tokai University

３．提案方式

３．１．　提案内容

現在のレコメンドシステムは「個人の情報」と「推奨対象の情報」という２つ情報で構成されたシステムであった。つまり２次元の情報で構成されたシステムであり情報の適合性が悪くなっているという現状がある。そこで今回の映画推奨システムでは「個人の情報」、「推奨対象の情報」の他に、「映画を見るときのシチュエーション」という第三の要素を取り入れ３次元な情報で構成されたレコメンドシステムの提案を行うと同時に、このシステムを階層的クラスタリング手法である最短距離法と最長距離法を使用した２つの映画推奨システムを作成し、それぞれの手法についての評価を行う。

３．２．　全体構成

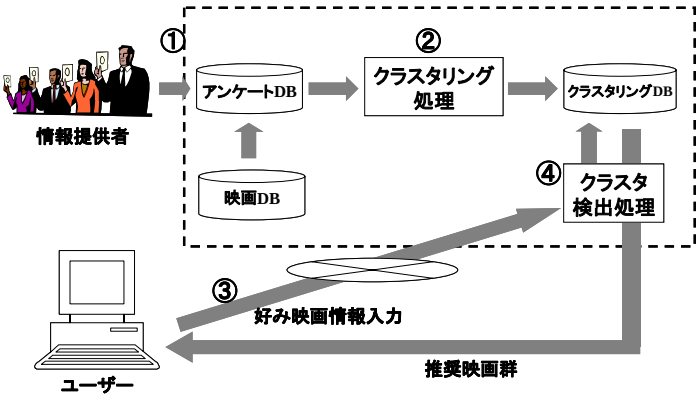

図１　全体構成図

図１は映画推奨システムの全体構成図である。以下に処理の説明をする。

①　アンケート情報収集

アンケート情報として情報提供者より「個人の情報」として好みのジャンル、年齢、職業、そして第三の要素の「映画を見るときのシチュエーション」として何時見たのか、何処で見たのか、誰と見たのかというこれらの情報を映画別に収集し、それらの情報をアンケートデータベースへ保存する。この時、「推奨対象の情報」としてそれぞれの映画のジャンルや監督等の情報は映画データベースより入手する。

② クラスタリング処理

アンケートデータベース内のデータにクラスタリングを行い分類分けを行いクラスタリングデータベースへクラスタリングされたデータを保存する。詳しい説明は３．３で説明する。

③ 好み映画情報入力

システムを利用するユーザーは「個人の情報」として年齢、職業、好みのジャンル、「推奨対象の情報」としてどんなジャンルが見たいのか、「映画を見るときのシチュエーション」として何時見るのか、何処で見るのか、誰と見るのかという７つの項目を入力する。

④ クラスタリング検出処理

③でユーザーより入力された情報から、適切なクラスタをクラスタリングデータベースより検出し、ユーザーへ検出したクラスタ内の映画情報群を出力する。

詳しくは下記の３．４で述べる。

３．３． クラスタリング処理

このシステムで使用するクラスタリング手法としては階層的クラスタリングの最短距離法、最長距離法を使用した２つのシステムを作成する。まず、アンケートデータベース内の映画情報を「個人の情報」、「推奨対象の情報」、「映画を見るときのシチュエーション」の３つのユークリッド空間へ配置する。そのとき「個人の情報」のユークリッド空間座標は年齢、好みのジャンル、職業という３次元空間で配置し、「推奨対象の情報」は、ジャンルの１次元空間、「映画を見るときのシチュエーション」は何時、何処、誰とという３次元空間座標で配置を行う。次にこの３つのユークリッド空間を合成し１つの合成ユークリッド空間を作成しこの空間で配置された要素で最短距離法、最長距離法を使用しそれぞれのアルゴリズムでクラスタリングを行う。

３．４． クラスタ検出処理

クラスタの検出処理としてはユーザーが入力した年齢、職業、好みのジャンル、見たいジャンル、何時、何処で、誰と見るのかという７つの情報を３．３で述べた３つのユークリッド空間、「個人の情報」、「推奨対象の情報」、「映画を見るときのシチュエーション」にそれぞれ配置しそれぞれの座標を合成し合成座標を算出する。そしてこの合成座標を３．３で述べた３つのユークリッド空間を合成した合成ユークリッド空間へ配置し、最も近い位置のクラスタを探し出し、ユーザーへ探し出したクラスタ内の映画情報群を出力するという処理を行う。

４． 評価

階層的クラスタリング手法の最短距離法と際長距離法、それぞれを使用した２つのシステムの比較・評価を行う。評価方法としては異なった趣向のユーザー２０人に３．２の③の好み映画情報入力を行ってもらい実際にそれぞれのシステムを使用してもらう。そのとき「個人の情報」はそれぞれのユーザーの情報を入力してもらい、全ての「対象物の情報」と「映画を見るときのシチュエーション」の組み合わせで行い、それぞれの組み合わせごとに出力された結果についての満足度を１０段階で算出してもらうという評価を行った。

５．考察・検討

本研究は階層的クラスタリングを用いた映画推奨システムの提案を行うと共に、同システムにおいて階層的クラスタリングのアルゴリズムの評価を行った。システムの考察としては、現システムでは最高３個の情報からなる３次元ユークリッド空間でのクラスタリングを行ったが、より的確に情報を絞り込む為には情報量を増やし、４次元、５次元という多次元ユークリッド空間での暮らす他リングを行う必要があると考える。又評価を行うクラスタリングのアルゴリズムも最短距離法や最長距離法だけでなく、様々なアルゴリズムの評価を行わなければならないと考える。

６．おわりに

本研究は映画推奨システムにおいてのクラスタリング手法である最短距離法と際長距離法の比較を行うことにより、映画推奨システムの実現を目指している。個人の趣向がますます多様化し、情報量が増加していく今日では、より的確に情報を推奨できるシステムが必要となってくる。その為、本研究のような情報推奨システムはこれから必要なシステムであると考える。

＜参考文献＞

[1] マイケル J. A. ベリー/ゴードン・リノフ、海文堂：「データマイニング手法」1999.

[2] 鷲尾泰俊/大橋靖雄:「多次元データ解析シリーズ入門 統計的方法３」 1989.

[3] 渡辺匠/太田学/片山薫/石川博：「文章間の差異に着目したクラスタリング手法の提案」情報処理学会第６７回全国大会／２E-２、2005年３月.

[4] 神嶌敏弘：http://www.kamisima.ney/jp/clustering/