5C-2

映画鑑賞の道案内システム

塩崎浩二　　浦谷則好
東京工芸大学　工学部　コンピュータ応用学科

１．はじめに

見たい映画を探すときに Web 検索をすることも多いが，既存の検索システムでは DVD 等の売れ筋情報をもとにしていることが多く，ジャンルや製作年代が同じものが推薦されることがほとんどである．また，Web サイトのレビュー記事などから内容を吟味して行くのは，その量の多さから非効率的である．本研究では，1 本の映画からあらすじのマッチングを取り，その映画の内容に関連が深い別の映画を推薦するシステムを提案する．このシステムを使用することで，ジャンルや製作年代に縛られず，次々と視聴したい映画のすそ野を広げていくことができる．

２．関連研究

あらすじのマッチングには連想検索エンジン GETA[1]を使用する．GETA は情報処理振興事業協会(IPA)が実施した「独創的情報技術育成事業」の研究成果であり，連想検索をはじめ，文書分類，単語間類似度計算など，大規模文書の分析を行うことができる．また，データベース形式は WAM という形式を使用する．WAM については後述する．また，館野らは GETA の単語間類似度計算を使用して，本との出会いを支援するシステム[2]を作成している．

３．映画推薦方法

3.1　映画の情報取得

推薦する映画の情報を Web サイト「allcinema」[3]「キネマ旬報映画データベース」[4](以下，キネマ旬報)の二つから取得する．まず「allcinema」から評価 8 段階のうち，推薦映画にふさわしい評価 5 以上の映画タイトル名を取得する．ここで得たタイトルを「キネマ旬報」で検索し，あらすじと製作年，監督名などの付帯情報を取得する．その際に，検索したタイトルに，数字がついたもの，「続」や「新」のついたものなどもシリーズものとして取得しておく．二つのサイトを併用したのは，「allcinema」はあらすじが充実しておらず，また「キネマ旬報」は既知のタイトル名から検索するシステムで，タイトルの一覧を得ることができないためである．

3.2　WAM ファイルの作成

得られたあらすじを形態素解析し，形態素ごとの出現回数をカウントし，複数回出現したものは一つにまとめる．また，固有名詞や助詞，助動詞，接続詞など，あらすじの意味合いに関係しないものを除く．それらを，GETA で使用する WAM 形式のファイルにする．

WAM は実際にコンピュータが計算に使用する内部表現と使用者が入力する外部表現があるが，ここでは外部表現について説明する．WAM は行指向の行列データファイルである．以下のように行(Row)と列(Column)を縦並びに書く．

	例
@タイトル 1	@暁の 7 人
3 特徴語 1	3 司令
2 特徴語 2	3 列車
1 特徴語 3	2 仲間
@タイトル 2	@アップルゲイツ
1 特徴語 4	5 破壊
5 特徴語 5	2 原子力
@タイトル 3	@暗黒街の特使
4 特徴語 6	4 ウィスキー
3 特徴語 7	3 密造

@に続けてタイトルを書き，その下にあらすじに出現する単語を書いていく．特徴語の先頭の数字は出現回数である．作成した外部表現の WAM を GETA の機能を使い内部表現のものに変換しておく．

3.3　付帯情報のファイル化

3.1 で得た付帯情報も，タイトル名とともに検索できるデータ形式でファイル化しておく．付帯情報は製作年，監督名，主演者名である．ここでは検索しやすく簡易な形式として，以下のように@タイトルの下に製作年，監督名，主演名を並べる．

The guidance system of movies
Koji Shiozaki, Noriyoshi Urtani
Tokyo Polytechnic University　Faculty of Engineering
Department of Applied Computer Science

@アップルゲイツ
1990 年
マイケル・レーマン
エド・ベグリー・ジュニア

3.4 インタフェース

インタフェースは Web ブラウザで表示できるように HTML 形式で記述し，マッチングプログラムを CGI で動作するようにする．ユーザが入力するワードは，お気に入りの 1 本の映画タイトル名，あるいは映画に関わったスタッフ名とする．検索のクエリとしては正確なタイトル名を前提としている．このため，「キネマ旬報」の検索システムを使用してタイトル名を確定する．スタッフ名の場合は関連する映画を選択決定し 1 つのタイトル名へ絞り込む．図 1 にシステムのフローを示す．

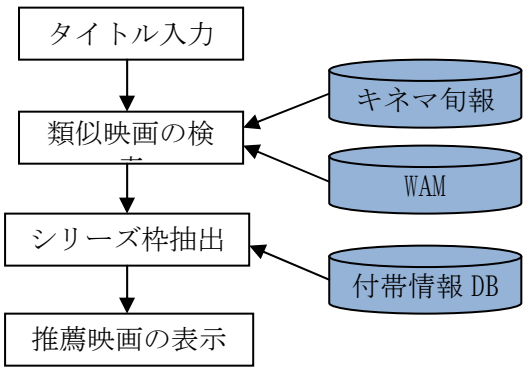

図 1　システムの概要

3.5 類似映画の検索

クエリとするタイトルが決まったら，その映画のあらすじを「キネマ旬報」から得る．リアルタイムであらすじを取得することで，クエリとする映画は最新のものにも対応することができる．そのあらすじを形態素解析し，得られた単語列と 3.2 で作成した WAM ファイルとの単語間の類似度計算を GETA で行う．類似度計算には tf（単語頻度）を使用する．その他に TFIDF 法や，それを拡張した SMART 法なども利用できるが，tf 法が最も的確な結果が得られている．類似度の高いものから 11 件のタイトルを取得する．11 件としたのは取得したい件数の 10 件の他に，自分自身が含まれるからである．

3.6 シリーズ枠の抽出

あらすじだけでは得られない関連性を得るために同シリーズ枠を設定する．たとえば主演者が同じものや，古い映画のリメーク版などである．シリーズ枠としては，同じシリーズのもの，同監督，同主演，同タイトルの４つを設定した．監督名や主演名を特定するために，3.3 でファイル化しておいた付帯情報を使用する．また，シリーズものはその中で直後に製作されたものが表示されるようにする．

3.7 推薦映画の表示

検索により得られた 10 件の映画（自分自身は除く）に対し，シリーズ枠に該当するものがある場合は，類似度が下位のものから最大で 4 件，シリーズ枠のタイトルに差し替え，計 10 件をお勧め映画として付帯情報とともに表示する．また，あらすじのリンクを作り，「キネマ旬報」サイトからあらすじを取得して表示させ，推薦された映画がどのような映画なのか，概要を見られるようにする．

４．評価

評価は，関連性があるとわかっているシリーズ枠の推薦映画を除いたものについて，「関連性がある(Excellent)」「関連性が少しある(Good)」「どちらとも言えない(Fare)」「関連性があまりない(Poor)」「関連性がまったく無い(Bad)」の 5 段階で行う．被験者 10 名程度にそれぞれ 5 件の映画を入力してもらい，1 件につき最大 10 本の映画が推薦されるので，最大計 500 本程度を評価対象と考えている．現在評価実験中である．

５．おわりに

本研究では映画情報取得の際に，Web サイトのページ送りが Javascript で記述されていたため，ソースコードを手動でコピーするなど半自動で行った．これを全自動にすることができれば，定期的に WAM ファイルを更新し最新の情報にすることの利便性が増す．また，WAM ファイルの内容が本システムの推薦結果を左右するので，作成には試行錯誤が必要である．たとえば，固有名詞をすべて除いてしまったが，フランスやニューヨークなどのような特徴的な地名を残すなど，より的確な単語の選別ができればよりよい推薦結果が得られると考えられる．

参考文献

[1]http://geta.ex.nii.ac.jp/geta.html
[2]館野紅理奈，浦谷則好：「本との出会い」を支援するシステム．言語処理学会第 17 回年次大会，P1-8，pp.190-193，2011
[3]http://www.allcinema.net/prog/index2.php
[4]http://www.kinejun.jp/