An application of social filtering to movie recommendation

D Fisk

The system described in this paper (MORSE — movie recommendation system) makes personalised film recommendations based on what is known about users' film preferences. These are provided to the system by users rating the films they have seen on a numeric scale. MORSE is based on the principle of social filtering. The accuracy of its recommendations improves as more people use the system and as more films are rated by individual users. MORSE is currently running on BT Laboratories' World Wide Web (WWW) server[1]. A full evaluation, described in this paper, was carried out after over 500 users had rated on average 70 films each. Also described are the motivation behind the development of MORSE, its algorithm, and how it compares and contrasts with related systems.

1. Introduction

We all have our own film preferences, and so take the risk, every time we watch a film, that we will not enjoy it. We can use a variety of methods to decide whether a film is worth watching, including the following:

- reading reviews,
- asking our friends (i.e. 'word of mouth'),
- finding out who starred in it or who directed it,
- consulting the cinema or video charts,
- some combination of the above.

All of these approaches have disadvantages. In the first two, the viewer is relying on somebody else's opinion about the film, which will not always agree with their own. The third method does not work reliably, as films from the same director can vary considerably in quality, and film stars can be miscast or give lack-lustre performances. The fourth does not indicate popularity so much as the rate of change of people's expectations. It measures the rate at which tickets are bought or videos hired, with no feedback indicating whether the viewer actually liked the films that they watched.

Of course, one may ask whether it is possible at all to predict anyone's likes or dislikes in any area where preferences are to a high degree subjective. Is there not truth in the old adages, 'there's no accounting for taste', and 'one man's meat is another man's poison'?

The answer, which has been shown in the results presented below, is that, to a certain extent, it is possible to predict preferences [1]. This is true because people's tastes are not unique and films are not all dissimilar.

The breakdown of the rest of the paper is as follows — section 2 explains the social filtering approach, its possible applications and its commercial exploitation, and summarises related systems. Section 3 describes the MORSE system's architecture and algorithm, section 4 presents the results of its evaluation, and in section 5, some alternative algorithms are evaluated. Section 6 concludes the paper.

2. Background and motivation

2.1 Social filtering

Social filtering is based on the 'word-of-mouth' approach to recommendation in that it relies on the opinions of others. However, it makes use of considerably more data than other word-of-mouth approaches typically do. The general method can be summarised as:

[1] At URL http://www.labs.bt.com/innovate/multimed/morse/morse.htm

- the user rates some films, CDs, etc,
- those ratings are compared with ratings given to the same films by other users,
- if you share similar tastes with others, it recommends to you those films that they liked, from among the films you have not rated.

As will be shown later, it is possible to improve the accuracy of the recommendations by taking into account ratings by people with dissimilar tastes too.

2.2 Possible applications

Similar techniques could be used to provide recommendations in other areas with similar properties, such as television programmes, personalised newspapers, music, novels, beers or wines. In common with films, these have similarities between items of both the domain (people) and range (the thing which is recommended). Also, in these areas, people cannot be sure about whether they like something until they try it. However, within each of these areas, the correlation between items is sufficient for information about one item to help in predicting a person's liking for other items. If these conditions are not met, content-based filtering (where a user specifies their likes and dislikes and the system checks items against these) could be used instead [2, 3].

2.3 Potential commercial benefits

Although, for the foreseeable future, a significant part of BT's revenue will come from narrowband services such as voice and data, it is expected that in the near future BT will provide broadband services to customers' homes by exploiting asymmetric digital subscriber line (ADSL) technology [4]. A substantial part of this bandwidth will be used for entertainment services, probably in the form of films and other video material. BT is already carrying out customer trials of video on demand [5].

As other companies (e.g. video shops, cable TV companies, telecommunications companies) will be competing with BT in the supply of such services, it is important that BT differentiates its services from those of the competition. There are two obvious ways of service differentiation — price and range. The scope for price cutting may be limited because of the costs of both launching and supplying such a service, and the intense competition expected in future markets. The problem with increasing the range of films is that it becomes impractical to browse around even a small fraction of a database in order to find an interesting film, and even then the selected film might be disappointing.

However, there is a third way — taking some of the chance out of the customer's film selection process. This can be achieved by making customised recommendations (by a method similar to the one outlined below) while at the same time providing objective information about how the recommendations were made. This will increase customer satisfaction by decreasing the likelihood that customers watch films they do not enjoy.

2.4 Related systems

Several systems have been developed which attempt to predict user preferences. Some of these are summarised in Table 1.

Table 1 Related user preference systems.

System	Domain	Comments
Movie Select [6]	Films	This is a CD ROM system claimed to incorporate artificial intelligence and fuzzy logic. It is accompanied by a comprehensive film database.
videos@ bellcore.com	Films	This is accessible by e-mail. It is unclear whether it uses social filtering [7], or is a neural network which makes use of film-specific information [8].
Movie Critic[a]	Films	This uses social filtering. Its algorithm is the subject of two patents.
Movie Recommen-dation Engine[b]	Films	No information on how this works is publicly available.
AgentMC[c]	Films	The developer "...hopes to get better performance [than Firefly, described below] by using multiple information sources and by adopting a well principled Bayesian approach"[9].
Ringo, HOMR [7]	Music	Developed at MIT. Ringo was accessible by e-mail, and was superseded by HOMR, and later Firefly.
Firefly[d]	Music, Films	This system was developed by the Autonomous Agents Group at MIT Media Labs, and is the successor to HOMR and Ringo. It uses social filtering.
The Similarities Engine[e]	Music	No information on how this works is publicly available.
GroupLens[f][10]	Usenet News	Developed at MIT Media Labs, this uses social filtering (Pearson-r algorithm).
NewT [11]	Usenet News	Also developed at MIT Media Labs, this uses genetic algorithms, with further non-genetic learning taking place between generations.
Yenta [12]	Match-making	Developed at MIT Labs, Yenta is a co-operating agents system which finds people with similar interests, clusters them and introduces them to each other.

a http://www.moviecritic.com/
b http://phoebe.dws.acs.cmu.edu/cgi-bin/movie
c http://HTTP.CS.Berkeley.EDU/~murphyk/AgentMC/AgentMC.html
d http://www.agents-inc.com/
e http://www.ari.net/se/
f http://www.cs.umn.edu/Research/GroupLens

3. Description and evaluation of MORSE

A large body of data was collected to evaluate the algorithm. In order to facilitate its use by as many people as possible and thereby improve its accuracy, MORSE had to satisfy the following two requirements:

- accessibility by as many people as possible,
- ease of use.

For this reason, it was decided to put MORSE on the World Wide Web, with data entered using forms. This had the additional advantages that the user interface (Netscape) would be familiar, so that instructions for users were unnecessary, and that it was straightforward to provide links into the Internet movie database, which contains comprehensive information about most films in the MORSE database.

A flow diagram for MORSE is given in Fig 1.

3.1 Description of MORSE's main modules

After registration (in which the user enters an e-mail address and password), the user selects from a menu which includes entering ratings and getting personalised recommendations, seeing a chart of films' average ratings, and some database management functions (see Fig 1). On choosing to enter ratings, the user is presented with a screen similar to Fig 2. The user can then rate films on an integer scale from 0 to 10, and repeat this on subsequent pages if desired.

After the user indicates to the system that the rating process is complete, recommendations are made by MORSE from among the films which remain unrated (see Fig 3). Both the predicted rating and the estimated error on it are taken into account when making recommendation. A film is:

- strongly recommended when its predicted rating, minus its estimated error, is at least 7,
- recommended when it is at least 6,
- weakly recommended when it is at least 5,
- not recommended if it falls below 5.

The user can obtain more information about the films by clicking on the films' titles. The relevant pages are then downloaded from the Internet movie database.

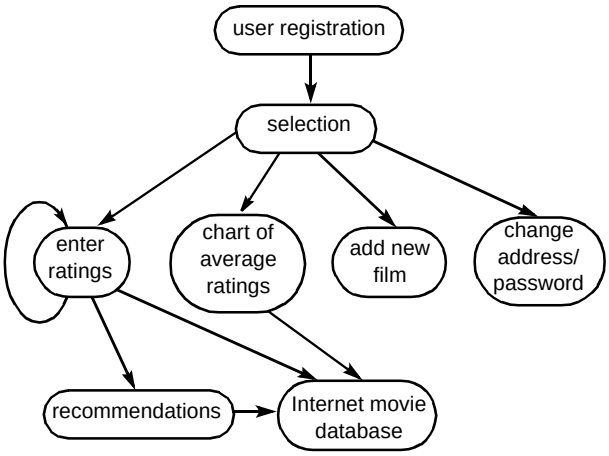


Fig 1 Flow diagram of MORSE.

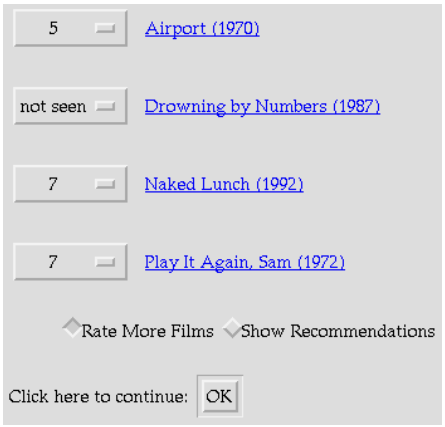


Fig 2 Entering ratings.

Predictions for anonymous:

	score	± error	count	mean	film
weakly recommended	9.21	± 4.05	21	8.10	Heavenly Creatures (1995)
recommended	8.78	± 2.06	71	7.90	The Remains of the Day (1993)
recommended	8.71	± 1.97	51	7.75	Chinatown (1974)
recommended	8.71	± 2.30	45	7.87	Hoop Dreams (1994)
strongly recommended	8.71	± 1.60	38	8.00	Lawrence of Arabia (1962)
strongly recommended	8.70	± 1.63	24	7.54	Heat (1995)

Fig 3 Recommendations.

3.2 The MORSE algorithm

The principle of the MORSE algorithm is as follows. The rating a user *i* would give to a film *k* is estimated by normalising the ratings other users gave to the same film, and then plotting them against the correlation between the users' ratings and extrapolating (see Fig 4) the best-fitting straight line to correlation = 1.0. The normalisation process involves plotting the user *i*'s ratings, for films whose ratings are closely correlated with those of film *k*, against those of each other user *j*, and using the best-fitting straight line to convert user *j*'s rating for *k* into an equivalent rating for *i* (see Fig 4).

The steps taken by the program are.

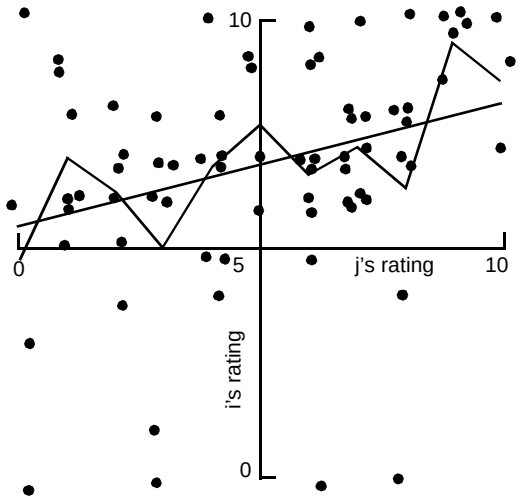
- *Step 1* — calculate the correlation between *k* and each other film.

For each other user *j*, do:

- *Step 2* — plot the ratings given by *i* to the *N* films most closely correlated with *k* against the ratings given to the same films by *j* (see Fig 4).
- *Step 3* — determine the best-fitting straight line through the points plotted in *Step 2* (see Fig 1). The best-fitting straight line is a function for converting the ratings given by *j* to *k* into equivalent ratings. Equivalent ratings are used because different users use the 0-10 scale differently.
- *Step 4* — determine the correlation between *i* and *j* for the *n* films most closely correlated with *k*.
- *Step 5* — plot the equivalent ratings for *k* (obtained in *Step 3*) against the correlation between *i* and *j* (obtained in *Step 4*) (see Fig 5).

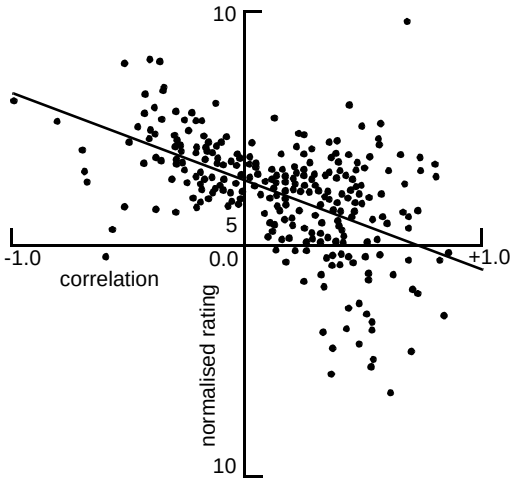
end for:

- *Step 6* — find the best-fitting straight line through the points plotted in *Step 5*. Where this crosses the line *correlation* = 1.0 is the estimated rating (see Fig 5). *Correlation* = 1.0 is equivalent to a hypothetical user who exactly shares the tastes of user *i*.
- *Step 7* — calculate the error on the extrapolation.
- *Step 8* — base recommmendations on the value obtained by subtracting the error (*Step 7*) from the estimated rating (*Step 6*). If results exceed a threshold, the film is recommended.



Plot of *i*'s rating against *j*'s rating for *n* films. The zigzag line is the average value of *i*'s ratings for each value of *j*'s ratings, and is included to show how it closely follows the best-fitting straight line.

Fig 4 *i*'s rating plotted against *j*'s rating.



Plot of each user's rating (normalised) against their correlation with *i*. *i*'s rating is estimated by extrapolating the best-fitting straight line to correlation = 1.0

Fig 5 Normalised ratings plotted against correlations.

4. Data analysis and results

The MORSE algorithm, and a number of alternatives, were evaluated by masking out the known ratings one by one, estimating each rating by using the algorithm, and then calculating the root mean square of the difference between each estimated rating and the actual rating given by the user.

At the time the results were evaluated, there were a total of 729 registered users, 562 of whom had rated at least one

film (users are not deleted if they do not enter ratings), and there were 675 films to rate. In all, there were 38 982 ratings entered, an average of 69 films per user. Unless stated otherwise, predictions were made for a total of 37 877 ratings in all. (Some of the films had been seen by fewer than ten people, and so their ratings were not predicted.)

4.1 Results

It was decided not only to measure the accuracy and certainty of MORSE's predictions, but also to determine the effect on accuracy of the number of correlated films used in the normalisation process, the number of users, and the number of films rated by individual users.

- The root mean squared (RMS) error was calculated to be 1.805 (between 1.792 and 1.818 with 95% confidence) for the MORSE algorithm — this is a significant improvement on the value 2.00 obtained when the average rating for the film was used as the prediction.

- 95.28% of recommendations (not predictions) were found to be satisfactory (here, a satisfactory recommendation is one where a film is recommended and the user enjoyed it) — more precisely, a recommendation is satisfactory if the film is strongly recommended and the user rated it 7 or above, or if it is recommended and the user rated it 6 or above, or it is weakly recommended and the user rated it 5 or above.

Because unsatisfactory recommendations would contribute to customer dissatisfaction, it is important to keep them to a minimum. Non-recommendations may have contained cases where a film would have been enjoyed by the user but was not recommended. This would not have contributed (significantly) to customer dissatisfaction, as the user would have wasted neither time nor money (the cost of video rental, for example) in not seeing the film.

- The correlation between the average predicted error for each user and the average RMS error was calculated to be + 0.43 (see Fig 6) — the correlation's positive value demonstrates that it is worth using the predicted error for determining the certainty of recommendations.

- The RMS error is a minimum when n, the number of correlated films used in the normalisation process, is equal to 11, though results are only significantly worse than this if n is outside the range 7 to 23.

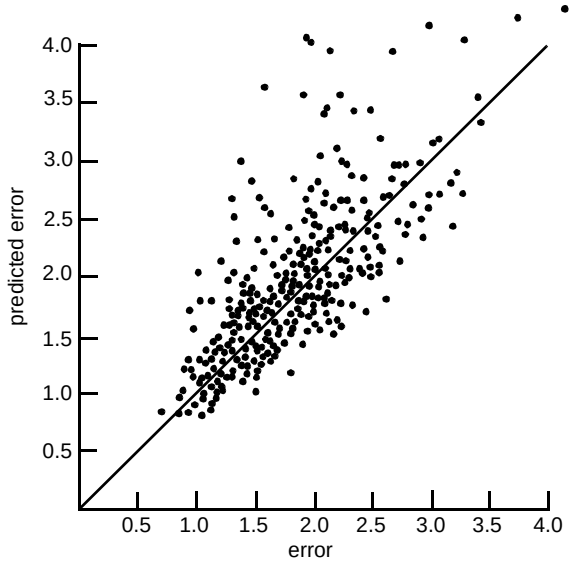


Fig 6 Accuracy of predicted errors.

The number of correlated films, n, used in the normalisation process (Fig 4) was varied, and the mean error was measured, in order to determine the value at which the error is a minimum. The results are shown in (Fig 7). If N is too low, the range of films used appears to be too small to be statistically significant, and, if n is too high, unrelated (and hence 'irrelevant') films are used in predicting the rating for a particular film.

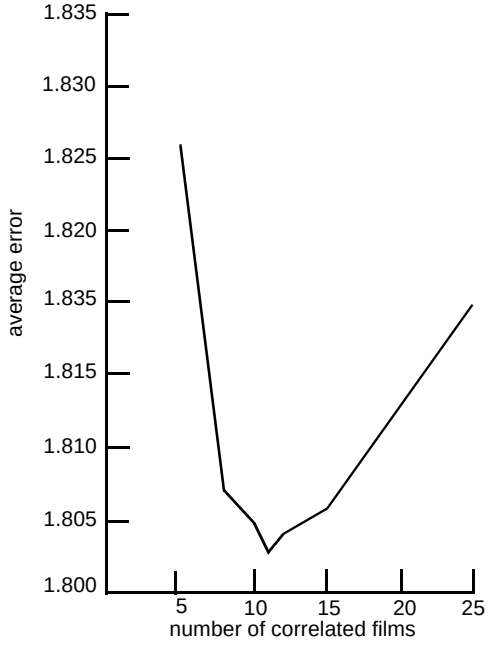


Fig 7 Variation of RMS error with n.

- The RMS error on predictions was found to decrease slowly, from over 2.2 to 1.8, as the number of users increased (see Fig 8).

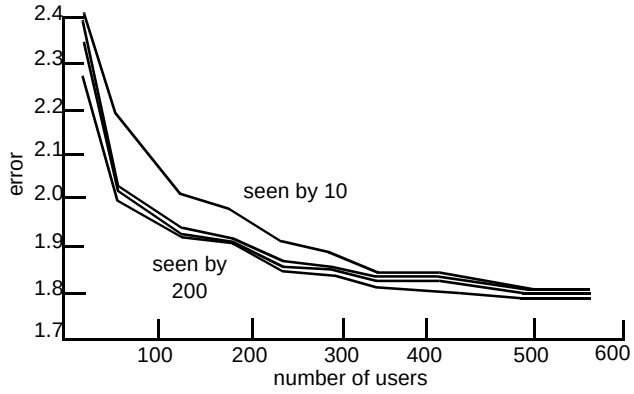


Fig 8 Variation of RMS error with number of users.

To explore the effect of increasing the number of users, fixed sets of films were rated (those which have been seen by fixed numbers of users (200, 100, 50 and 10) taken from the full data set). Then, the performance of MORSE was investigated as the number of users was varied (by deleting randomly selected users from the database). The results are plotted in Fig 8. They show a steady improvement as the number of users increases, and a slight improvement as the film range is progressively narrowed to the films rated by the most people. Intuitively, this was expected, since the quantity of information on which the predictions are based increases as the user base increases. In particular, the number of users with very similar (and also very dissimilar) tastes to any particular user increases, resulting in a more accurate extrapolation of the best-fitting straight line (as previously illustrated in Fig 5).

- The RMS error for a user was found to decrease slowly, from over 2.2 to under 1.6, as the number of films seen by the user increased (see Fig 9).

The results show that, for users to get accurate recommendations (RMS error < 2.0), they should rate at least 30 films, and that predictions improve in accuracy as more films are rated. The reason for this might be that, when predicting the rating for a given film, the *n* films used to calculate the normalised ratings (Fig 4 above) are more closely correlated with it. Also, the correlations between users can be more accurately determined, resulting in more accurate extrapolation (as previously shown in Fig 4).

The line through the points on Fig 9 was obtained by grouping the points into sets of 50 with adjacent values of the *x*-co-ordinate (the number of films rated), and connecting the average values of the *x*-co-ordinate and the *y*-co-ordinate (the average error) in each set of points.

- The mean intrinsic error was found to be 1.535 (between 1.466 and 1.610 with 95% confidence).

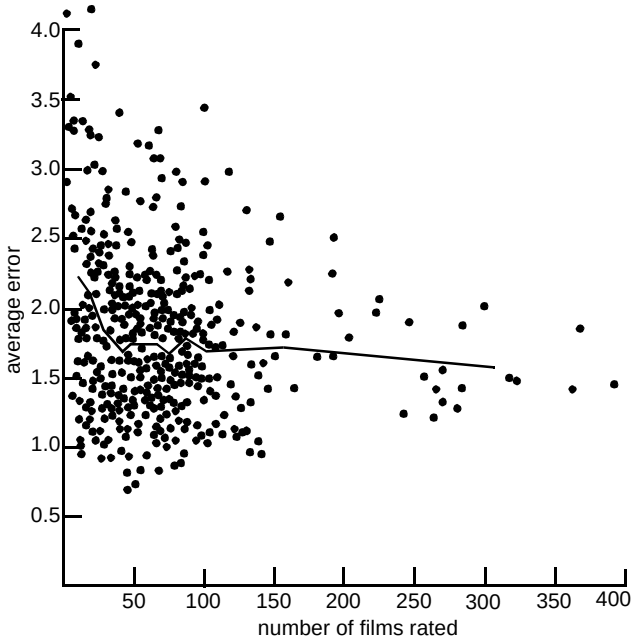


Fig 9 Variation of RMS error with number of films rated.

There are several reasons for people varying the rating they give to a film. The first is that, as they grow older, their tastes change (for most people, one would expect a secular decrease in their rating of a Disney cartoon, for example). The second is a tendency to give more recently released or viewed films a higher rating than the same film would have if it was less recent. A perusal of the top 100 films in the Internet movie database [1] will confirm this. The third (the intrinsic error or variance) is essentially random and depends on how consistently or carefully the user rates a film. Even though a user may systematically rate films higher or lower on different occasions, the amount by which they do is still likely to be random.

The intrinsic error is an absolute limit on the accuracy of MORSE or any similar system — no system could generate more accurate results than the user who originally produced the ratings.

This third source of error was estimated by asking 14 users to rate the first 100 films in the database a second time, without them referring to the ratings they originally gave to the list of films. Although these users were not picked at random (it would have been difficult to pick volunteers at random), they were not selected by any other criterion than their availability. Altogether, 882 ratings were used.

[1] http://us.imbd.com/

4.2 Filtering out spurious data using Chauvenet's criterion

A slight modification of the algorithm, intended to remove possibly spurious data, was attempted. This involves the use of Chauvenet's criterion.

Chauvenet's criterion [13] assumes that data points follow a Gaussian distribution. The fewer points there are, and the greater the distance from the centre of the distribution, the less likely it will be that any points will be found beyond that distance. If the probability of finding a single point beyond a certain distance is less than a fixed value (usually taken to be 0.5), any points which do lie beyond it can be disregarded as they are likely to:

- be spurious,
- disproportionately influence any deductions which are based on the distribution, e.g. the intercept and slope of the best-fitting straight line.

The best-fitting straight line, and the error on this, is calculated. After points are removed using Chauvenet's criterion, the best-fitting straight line, and hence the predicted rating, is recalculated.

The result was a slight improvement, but not a significant one, on the basic MORSE algorithm.

The RMS error was 1.801 (between 1.788 and 1.814 with 95% confidence). If it is a real effect, it may have been because a few users were careless when entering ratings, or maliciously entered false ratings, or have such idiosyncratic tastes that their ratings cannot be used to predict other people's. When n was increased from 10 to 11, the use of Chauvenet's criterion gave the same result.

5. Alternative algorithms

There are potentially an infinite number of alternative algorithms. These can be categorised into trivial algorithms and more complex ones.

5.1 Trivial algorithms

Three trivial algorithms were tried because, while they are not useful for predicting users' ratings, they provide a baseline against which the performance of more complex algorithms can be measured.

- Using the average rating given by the user as the prediction, the RMS error on predictions was found to be 2.072 (between 2.057 and 2.087 with 95% confidence).

  This is equivalent to providing the user with no information, because users can calculate this by averaging their ratings.

- Using the average rating given to each film as the prediction, the RMS error was found to be 2.000 (between 1.986 and 2.014 with 95% confidence).

  This is an improvement on the previous figure of 2.072 and suggests that there is a certain amount of consensus as to which films are good, and which are bad.

- Using the nearest neighbour's rating for each film as the prediction, the RMS error was found to be 4.177 (between 4.147 and 4.207 with 95% confidence).

  Each rating is estimated to be the rating given by the user who has seen the film and who has the closest tastes to the user whose rating is being estimated. That this method is so inaccurate is surprising, especially since it is similar to what so many people do in practice, namely to ask one of their friends for recommendations.

5.2 The Pearson-r algorithm

Using the Pearson-r algorithm, the RMS error was 1.940 (between 1.897 and 1.984 with 95% confidence).

Pearson-r is the algorithm used in GroupLens [8]. This algorithm ran considerably slower than the MORSE algorithm, probably because the data structures used were designed for the MORSE algorithm rather than the Pearson-r algorithm. Because of the slow processing speed, it was decided to run the algorithm on a randomly selected 10% sample of the data set (3826 ratings). The simultaneously calculated RMS error for the MORSE algorithm was 1.793 (between 1.754 and 1.834 with 95% confidence). It is 2.018 (between 1.974 and 2.064 with 95% confidence), if the average is used as the predicted rating. Thus, the MORSE algorithm makes more accurate predictions than the Pearson-r algorithm when using the same data.

5.3 Average of M nearest neighbours

The lowest RMS error obtained using the average of M nearest neighbours was 1.926 (between 1.912 and 1.940 with 95% confidence).

To predict user i's rating for film k, the films are sorted in order of their correlation with k, and the closest n selected. Then, the ratings given by the M users most closely correlated with i (using only the closest n films to calculate the correlation between i and other users), who have seen film k, are averaged. The average value is the predicted rating for user i and film k.

The prediction improves as n increases, and is best if all the films are used in the comparison (n = maximum). This is different from the results using the MORSE algorithm, where the best results were obtained with $n = 11$ (see Fig 7). The reason for this is unknown.

If $M = 1$ (i.e. only the nearest neighbour is used), the prediction is very poor. It should be noted that this is equivalent to asking the person, whose film tastes most closely match your own, what they thought of the film, and is a frequently adopted strategy. If all users are employed in the comparison, the predicted rating is the average rating (and the error on this is 2.000). Values of M close to 25 are, however, improvements on this. The error minimised at 1.926 (between 1.912 and 1.940 with 95% confidence), a significant improvement on using the average as the prediction, but not as good as MORSE (Fig 10).

The implication of this is that film recommendations made by one friend (e.g. the one with tastes most similar to your own) is not a good way of deciding which film to watch. It is better to ask all your friends (preferably at least 20 of them) before making a decision (better still, use MORSE). This may be because we are more similar to an average of all our friends than to any individual friend. The same principle would apply even if the 'friends' are anonymous, as is typically the case in systems which employ social filtering.

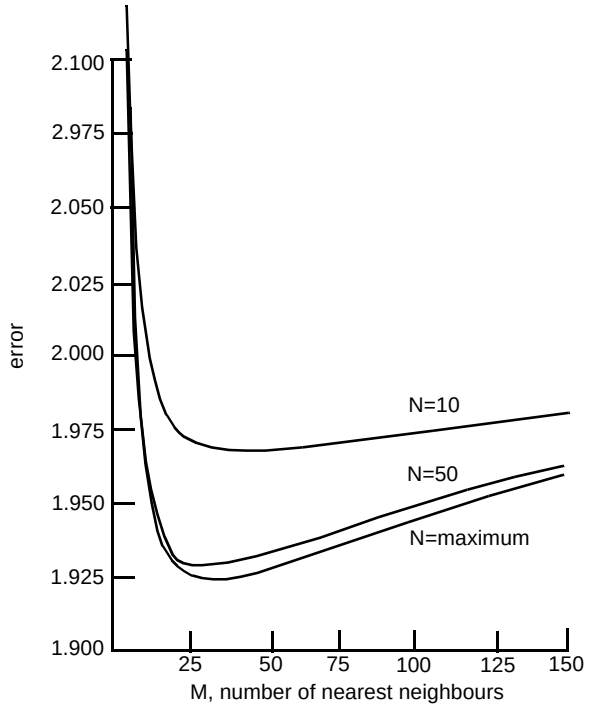


Fig 10 Error plotted against number of nearest neighbours.

6. Conclusions

The research has confirmed that it is possible to provide personalised film recommendations using only the ratings given by users for those films. The recommendations obtained by applying the MORSE algorithm were more accurate than the other algorithms tried. This confirms that it is possible and worthwhile to predict people's subjective tastes. It should be possible to use similar techniques in other areas where personal tastes determine user preferences.

Acknowledgement

Thanks are due to Richard Titmuss for helping with some of the low-level C Unix hacking, to data mining colleagues for advice on statistical techniques, and to Hyacinth Nwana and Robin Smith for suggesting improvements to the paper.

References

1 Fisk D: 'Recommending Films Using Social Filtering', BT MSc Dissertation (1995).

2 Information Filtering Bibliography — http://ils.unc.edu/gants/filterbib.html

3 Information Filtering Resources — http://www.enee.umd.edu/medlab/filter/filter.html

4 Young G, Foster K T and Cook J W: 'Broadband delivery over copper', BT Technol J, 13, No 4, pp78—96 (October 1995).

5 Whyte W S: 'The many dimensions of multimedia communications', BT Technol J, 13, No 4, pp 9—20 (October 1995).

6 Medior Incorporated: 'Movie Select: The Intelligent Guide to Over 44,000 Videos', (CD ROM) Paramount Interactive (1995).

7 Shardanand U: 'Social Information Filtering for Music Recommendation', MIT Media Laboratory Learning and Common Sense Group Technical Report 94-04 (1994).

8 Karunanithi N and Alspector J: 'A Feature-Based Neural Network Movie Selection Approach', Proceedings of the International Workshop on Applications of Neural Networks to Telecommunications, Stockholm (1995).

9 Murphy K P: 'CS289 Final Project Proposal: A Bayesian Movie Critic', http://HTTP.CS.Berkeley.EDU/~murphyk/AgentMC/proposal.ps

10 Resnick P, Iacovou N, Suchak M, Bergstrom P and Riedl J: 'GroupLens: An Open Architecture for Collaborative Filtering of Netnews', Proceedings of the ACM Conference on Computer Supported Cooperative Work, Chapel Hill, NC, pp 175—186 (1994).

11 Maes P: 'Evolving Agents for Personalized Information Filtering', Proceedings of the Ninth Conference on AI for Application, CAIA '93 (1993).

12 Foner L: 'A Multi-Agent Referral System for Matchmaking', Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM '96), London (April 1996).

13 Barford N C: 'Experimental Measurements: Precision, Error and Truth', p102, Addison Wesley (1967).

14 Fisk D: 'Programme Transmission and Reception System', EP Application No. 95305539.9 (BT Case Reference A25026) (1996).

Donald Fisk received a BSc (Hons) in Physics and Astronomy from Glasgow University in 1978, then did some research on General Relativity at Queen Mary and Westfield College, before his first job in virtual machine maintenance/development at Burroughs Machines Ltd in Cumbernauld. In 1984 he emigrated to Hong Kong where he worked on Compiler Development (Hong Kong Poly), Expert Systems (TI) and Speech Processing (HK Productivity Council), before returning to the UK to work for BT in advanced information processing. He has recently completed the BT MSc, for which he developed MORSE, a WWW-based movie recommendation system.