2C5-2

推薦結果の意外性を評価する指標の提案

Metrics for evaluation of Serendipity of recommendation list

村上 知子　森 紘一郎　折原 良平
Tomoko Murakami　Koichiro Mori　Ryohei Orihara

株式会社東芝 研究開発センター
Corporate R&D Center, Toshiba corporation

Recommender systems have been evaluated in many ways. Although prediction quality is frequently measured by various accuracy metrics, recommender systems must provide not only accuracy but also usefullness. A few researchers have argued that the bottom-line measure of recommender system success should be user satisfaction. In this paper we propose metrics *unexpectedness* and *unexpectedness_r* to measure serendipity of recommendation lists by recommender systems. The basic idea of the proposed metrics is that unexpectedness is the distance between the result by the method to be evaluated and that by primitive predicition method. *unexpectedness* is a metric for a whole recommendation list whereas *unexpectedness_r* is that taking account of ranking in the list. From the both point of accuracy and serendipity, we evaluate the results by three prediction methods in the experimental studies of television program recommendation.

1. はじめに

情報推薦において，推薦システムが利用者に有益な情報を提供することは最も重要な課題のひとつである．利用者の嗜好に適した推薦結果を提示することが推薦システムの有用性を決定づけるという観点から，情報検索の分野で確立された評価指標を適用して，予測の正確さ (accuracy) が評価されることが多い．推薦システムの有用性は，利用者の嗜好に適した推薦結果を提示することに加え，利用者に未知の情報や意外な情報を提供することによる驚きや喜びを提供することにも大きく関係している．情報検索の分野で確立された評価指標が利用者の嗜好に対する推薦結果の適合度合いを評価する指標であるのに対して，利用者にとって推薦結果が未知な情報や意外な情報である度合いを計る指標の研究は，情報推薦における未着手の課題である．

本論文では，推薦結果の意外性を評価する指標を提案する．推薦リストの意外性は，その推薦リストを生成した予測方法の予測結果とプリミティブな予測方法による予測結果との差異にあると仮定し，*unexpectedness* と *unexpectedness_r* の 2 種類の評価指標を提案する．そして，提案する指標を用いて推薦リストの意外性を計算することにより，予測手法を評価する．テレビの放送番組の推薦実験を通じて，予測の正確さと推薦結果の意外性を評価し結果を考察する．

本論文は以下のように構成される．2 章で推薦結果の評価指標に関して従来研究を紹介する．3 章で推薦結果の意外性を評価する提案指標に関して述べる．4 章でテレビ番組の推薦における予測手法の評価実験に関して説明する．5 章で結論と今後の展望を述べる．

2. 推薦結果の評価指標

従来，利用者の嗜好に適した推薦結果を提示することが推薦システムの有用性を決定づけるという観点から推薦システムが評価された [Breese 98, Billsus 98, Sarwar 00]．主な評価指標として，予測誤算の考え方を適用した MAE(Mean Absolute Error)[Breese 98]，情報検索の分野で確立された precision and recall [Cleverdon 68]，F-measure [Sarwar 00]，ROC(Relative Operating Characteristic)[Swets 69] などが挙げられる．また，予測順位と利用者による評価順位との相関を計算する相関係数 [Hill 95]，予測順位が上位であることの価値を指数関数的に評価する Half life utility metric [Breese 98] なども利用される．これらの指標を適用することにより，推薦システムが利用者の嗜好に適合するアイテムを推薦しているかどうか，もしくは利用者の嗜好に適合するアイテムを推薦リストの上位で推薦しているかどうかを評価することができる．

推薦結果が利用者の嗜好に適していれば，推薦システムに対する信頼度は高まるが，永続的な推薦システムの利用のためには，利用者に対して既知の情報のみならず，未知の情報や意外な情報の推薦も重要である [Swearingen 01]．

利用者に新たな価値を提供する推薦システムの重要性は，過去の研究者も言及している [Sarwar 01, Herlocker 04, Ziegler 05]．Sarwar らは，協調フィルタリングによる推薦システムにおいて，新規性や意外性のある推薦リストの作成を検討した [Sarwar 01]．彼らは，多くの利用者に人気のアイテムが推薦される傾向にある協調フィルタリングのアルゴリズムを，対象利用者が好むアイテムの推薦を優先するように改良し，意外性を高める工夫をした．Herlocker らは，対象利用者の嗜好とコミュニティ(多くの利用者) の嗜好を両方利用することにより意外性の高い推薦リストが作成可能であるとしている [Herlocker 04]．たとえば，(対象利用者が好む確率/コミュニティのメンバーが好む確率) を定義し，その値を基に推薦リストを作成することによって，推薦リストの内容が劇的に変化し，対象利用者が好む人気のないアイテムを推薦することができれば，推薦結果の意外性は高くなるであろうと推測している．Ziegler らは，類似するアイテムよりも多様な分野のアイテムで構成される推薦リストの方が利用者に対して有益であると仮定し，推薦結果の多様性 (Diversity) を評価する指標を提案した [Ziegler 05]．さらに，多様な分野のアイテムを含む推薦リストを作成するアルゴリズムを提案し，precision や recall は低下するものの多様性は向上したことを本の推薦の実験で示した．

Herlocker らが推察するように，たとえ対象利用者が好む人

連絡先: 村上 知子，(株) 東芝研究開発センター，
住所:〒 212-8582 川崎市幸区小向東芝町 1
電話番号:044-549-2406
メイルアドレス:tomoko.murakami@toshiba.co.jp

気のないアイテムを推薦することができたとしても，それらのアイテムが利用者の習慣的な行動の対象であれば，利用者に対する推薦結果の意外性は高くないと予想する．例えば，テレビ番組の推薦の例で言えば，比較的視聴率の低い「囲碁・将棋」の番組を習慣的に視聴している利用者に対して，それらの番組を推薦したとしても，その推薦結果は意外ではないはずである．視聴者が習慣的に視聴している毎週放送の番組を推薦システムが数多く推薦する場合には，推薦結果の適合率は高いが，利用者が推薦によって新たな番組の存在を発見することは難しい．また，Ziegler らは，利用者の嗜好に対する適合性よりも推薦結果の意外性を重要視しているが，推薦システムに対する利用者の信頼度と永続的な利用の観点から，両方を評価する評価指標が必要であると考える．

3. 提案指標

本論文では推薦結果の意外性を評価するための指標を提案する．提案する指標による基本的な評価方式の説明図を図 1 に示す．図 1 に示すように，推薦結果の意外性は，評価対象である推薦システムの予測結果とプリミティブな推薦システムの予測結果との差異にあると仮定し，それらを比較することによって求めることとした．つまり，プリミティブな予測方法でも予測が可能なアイテムは意外性が低く，逆にプリミティブな予測方法では予測が困難なアイテムに意外性が高いと考えた．プリミティブな予測方法とは，容易に予測が可能な基本的な予測方法のことを指し，利用者プロファイルや過去の利用者の行動履歴情報に基づいて推薦する方法などが考えられる．

L 件の推薦リストの意外性を評価する指標として $unexpectedness$ を提案する．推薦リスト中第 i 位の推薦アイテムを s_i，評価対象の予測方法による s_i の予測確率値を $P(s_i)$ とする．プリミティブな予測方法による s_i の予測確率値を $prim(s_i)$，利用者の興味に対する s_i の適合度合いを $isrel(s_i)$ とする．$unexpectedness$ を以下の式を用いて計算する．

$$unexpectedness = \frac{1}{L}\sum_{i=1}^{L}\max(P(s_i)-prim(s_i),0)\cdot isrel(s_i) \quad (1)$$

評価対象の予測方法による結果とプリミティブな予測方法による結果との差異を式 (1) の max 項で計算し，$isrel$ 項を掛けることによって，利用者の嗜好に適合しかつ意外なアイテムのみを計算の対象とするようにした．

式 (1) では，意外性のあるアイテムの順位を考慮していない．すなわち，同件数の同予測確率値のアイテムが含まれる推薦リストを式 (1) を用いて評価すると，それらが推薦リスト中の上位に存在しても下位に存在しても同じ評価値が出力される．そこで，意外性のあるアイテムが推薦リストの下位よりも上位に存在する推薦リストを高く評価する指標，$unexpectedness\_r$ を提案する．上位 i 件以上の適合アイテム件数を $count(i)$ として，$unexpectedness\_r$ を以下の式を用いて計算する．

$$unexpectedness\_r = \frac{1}{L}\sum_{i=1}^{L}\max(P(s_i)-prim(s_i),0)\cdot isrel(s_i)\cdot\frac{count(i)}{i} \quad (2)$$

式 (2) は，式 (1) に $\frac{count(i)}{i}$ の項の掛け算を適用することで順位を考慮した評価指標になっている．

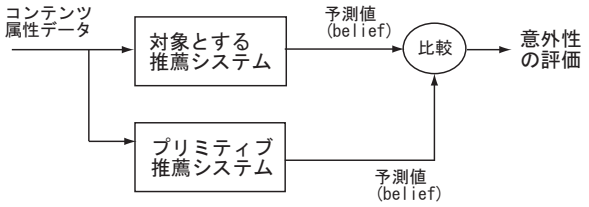

図 1: 評価方式

4. 推薦結果の評価

テレビの放送番組を対象とした推薦の実験を通じて，3 種類の推薦手法による推薦結果を評価した．

4.1 データ

実験に必要な放送番組のデータは，電子番組表 (Electronic Program Guide) から収集した．放送番組のデータは，1 日あたり約 800 件の地上波/BS 放送の番組で構成されている．実験は一般視聴者 46 人を対象として，アンケート調査によって実際に視聴・録画した番組，もしくは実際には視聴しなかったが視聴したかった番組を，利用者の嗜好に関する「視聴・録画データ」として網羅的に収集した．調査期間は 2 回に分けて，2007 年 2 月 3 日から 16 日までの 2 週間と，2007 年 3 月 3 日から 9 日までの 1 週間とした．また，アンケート調査の自由記述形式にて各視聴者の好きな出演者情報を収集，103 種類の番組ジャンル情報からの任意数の選択形式にて好きなジャンル情報を収集した．

4.2 実験方法

調査期間前半の 2 週間に収集した EPG データと視聴・録画データを基に視聴の予測モデルを学習し，後半の 1 週間の調査期間は EPG データを用いて 1 日 1 回各視聴者に対して推薦リストを配信した．推薦リストは，放送番組に対する視聴者の視聴を予測することによって作成される．本実験では，以下の 3 種類の予測方法によって視聴者の視聴を予測し，推薦リストを作成した．

BN1 番組ジャンルと出演者に対する嗜好から視聴を予測するベイジアンネットモデルに基づいて視聴確率を計算し，推薦リストを作成する方法

BN2 番組ジャンルと出演者に対する嗜好と習慣性から視聴を予測するベイジアンネットモデルに基づいて視聴確率を計算し，推薦リストを作成する方法

KF EPG データの番組内容に出現するキーワードに重み付けをして，それらを基に視聴確率を計算し，推薦リストを作成する方法

Breese らが示したように，ベイジアンネットモデル [Pearl 88] による予測の精度は高いことで知られているが [Breese 98]，予測の精度はそのモデルの構造に依存するため，本実験では構造の異なる BN1 と BN2 の 2 種類のモデルを導入した．KF にはスパムメイルフィルタリングで効果のある Graham らの手法 [Graham 02] を適用した．これは，最も視聴・録画番組数の多い視聴者でさえも，全番組に対する視聴・録画番組数の割合が 2.5%に過ぎないことから，視聴の予測問題を大量のスパ

ムメイルに対して非常に少ないハムメイルを選別する問題と同様と見なしたためである．

3 種類の予測手法による推薦結果を，precision，recall，unexpectedness(式 (1))，unexpectedness_r(式 (2)) を用いて評価した．precision は推薦リストに占める適合アイテム件数の割合，recall は全適合アイテムに占める推薦リストに含まれるアイテム件数の割合であるから，全適合アイテム数を R とすると，L や $count(i)$ を利用して以下の式で求められる．

$$precision = \frac{count(L)}{L} \quad (3)$$

$$recall = \frac{count(L)}{R} \quad (4)$$

unexpectedness，unexpectedness_r の計算で利用する，プリミティブな予測手法には，アンケート調査で収集した視聴者の好きな出演者，好きな番組ジャンル，前半の 2 週間の視聴調査期間における視聴時間枠とのマッチングにより推薦番組を決定する手法を採用した．そのため，$prim(s_i)$ は 0 または 1 の 2 値をとる．また，視聴者の興味に対する番組の適合度合いは，視聴・録画データを利用し，実際に視聴・録画した番組，あるいは実際には視聴しなかったが視聴したかった番組に対しては $isrel(s_i) = 1$，それ以外に対しては $isrel(s_i) = 0$ とした．

4.3 評価

推薦リストのアイテム件数が 20 件 (L=20) の場合の評価値を表 1 に示す．表 1 は，BN1，BN2，KF の予測方法に対する 4 つの評価指標による評価値に関して，実験日ごとに全視聴者の平均値を求めた結果である．表 1 の 1 段目に，3 つの予測方法に対する precision と recall による評価値を示す．視聴者の好きな番組ジャンル，好きな出演者，過去の視聴時間枠とのマッチングに基づく 3 種類のプリミティブな予測に対する評価結果を，それぞれ表の 2 段目 (primitive=genre)，3 段目 (primitive=celebrity) 4 段目 (primitive=timeframe) に示す．表 1 から，他予測手法と比較して，BN2 が precision と recall ともに最も高い精度を示した．BN2 が BN1 よりも高い値を示していることから，視聴の予測精度の向上に習慣性の導入による効果があることがうかがえる．また，いずれのプリミティブな予測方法においても，他の予測方法と比較して KF が *unexpectedness*，*unexpectedness_r* に関して高い値を示した．これらの結果から，提案した評価指標を用いると，3 種類の予測方法の中では KF が精度を維持しつつ最も意外性が高い推薦結果を提示する予測方法であることが分かる．アンケート調査で実際に視聴者に対して推薦番組の意外性を質問した結果と比較したところ，3 種類の予測手法による推薦リストに含まれる意外な番組数は BN1，BN2，KF でそれぞれ 17，173，194 件 (全視聴者合計) であることからも，評価指標による結果が妥当であること明らかになった．

5. おわりに

本論文では，推薦結果の意外性は，評価対象である推薦システムの予測結果とプリミティブな推薦システムの予測結果との差異にあると仮定し，*unexpectedness* と *unexpectedness_r* の 2 種類の評価指標を提案した．さらに，テレビの放送番組の推薦実験を通じて，予測の正確さと推薦結果の意外性を評価し結果を考察した．今後は，視聴者にとって実際に意外な番組のアンケート調査データ基づいて，提案指標の妥当性，3 種類のプリミティブな予測方法の妥当性を比較，検証していく．

参考文献

[Breese 98] J. Breese, J. Herlocker and C. Kadie.: Empirical analysis of predictive algorithms for collaborative filtering, In proc. of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98), pp.43-52, (1998).

[Cleverdon 68] C. Cleverdon and M. Kean: Factors Determining the Performance of Indexing Systems, Aslib Cranfield Research Project, Cranfield, England. (1968).

[Billsus 98] D. Billsus and M. Pazzani: Learning collaborative information filters, In proc. of the 15th National Conference on Artificial Intelligence(AAAI), pp.46-53, (1998).

[Sarwar 00] B. Sarwar et al.: Analysis of recommendation algorithms for E-commerce, In proc. of the 2nd ACM Conference on Electronic Commerce (EC00), pp.285-295, (2000).

[Swets 69] J. Swets: Effectiveness of information retrieval methods, Amer. Doc. 20, pp.72-89, (1969).

[Hill 95] W. Hill et al.: Recommending and evaluating choices in a virtual community of use, in Proc. of Conference on Human Factors in Computing Systems(ACM CHI), pp.194-201, (1995).

[Swearingen 01] K. Swearingen and R. Sinha: Beyond Algorithms: An HCI Perspective on Recommender Systems, ACM SIGIR Workshop on Recommender Systems, (2001).

[Sarwar 01] M. Sarwar, G. Karypis, A. J. Konstan and J. Riedl: Item-based collaborative filtering recommendation algorithms, In proc. of WWW'01, pp.22-32, (2001).

[Herlocker 04] J. Herlocker, J. Konstan, L. Terveen and J. Riedl: Evaluating Collaborative Filtering Recommender Systems, J. of ACM Transactions on Information Systems, Vol.22, No.1, pp.5-53, (2004).

[Basu 98] C. Basu, H. Hirsh and W. Cohen: Recommendation as Classification: Using Social and Content-Based Information in Recommendation, In proc. of AAAI, pp.714-720, (1998).

[Ziegler 05] Cai-Nicolas Ziegler, S. M. Mcnee, J. A. Konstan and G. Lausen: Improving Recommendation Lists Through Topic Diversification, In proc. of WWW'05, pp.22-32, (2005).

[Pearl 88] J. Pearl: Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, CA, (1988).

[Graham 02] P. Graham: A plan for spam(http://www.paulgraham.com/spam.html), August, (2002).

表 1: 推薦リストの評価 (L=20)

	BN1		BN2		KF	
date	precision	recall	precision	recall	precision	recall
2007/3/3	0.2630	0.3583	0.4435	0.6075	0.3283	0.4524
2007/3/4	0.2087	0.2552	0.4217	0.5661	0.2793	0.3495
2007/3/5	0.3707	0.4455	0.4978	0.6201	0.3783	0.4477
2007/3/6	0.3446	0.4231	0.4924	0.6252	0.3935	0.4731
2007/3/7	0.3609	0.4509	0.4793	0.6074	0.4261	0.5037
2007/3/8	0.3565	0.4369	0.5098	0.6275	0.4065	0.4764
2007/3/9	0.3457	0.4270	0.4848	0.6019	0.4152	0.4819
average	0.3214	0.3996	0.4756	0.6080	0.3753	0.4550
primitive=genre						
	unexpected	unexpected_r	unexpected	unexpected_r	unexpected	unexpected_r
2007/3/3	0.0431	0.0236	0.1372	0.1018	0.1706	0.0921
2007/3/4	0.0304	0.0121	0.1088	0.0762	0.1422	0.0811
2007/3/5	0.0694	0.0416	0.1325	0.0980	0.1989	0.1285
2007/3/6	0.0598	0.0351	0.1273	0.0904	0.2000	0.1187
2007/3/7	0.0668	0.0414	0.1365	0.0979	0.2283	0.1393
2007/3/8	0.0622	0.0401	0.1451	0.1061	0.2022	0.1138
2007/3/9	0.0589	0.0332	0.1188	0.0797	0.2022	0.1241
average	0.0558	0.0324	0.1295	0.0929	0.1921	0.1140
primitive=celebrity						
	unexpected	unexpected_r	unexpected	unexpected_r	unexpected	unexpected_r
2007/3/3	0.0774	0.0406	0.2273	0.1606	0.3097	0.1676
2007/3/4	0.0621	0.0273	0.2265	0.1616	0.2726	0.1533
2007/3/5	0.1225	0.0707	0.2598	0.1886	0.3674	0.2217
2007/3/6	0.1184	0.0667	0.2663	0.1896	0.3804	0.2210
2007/3/7	0.1212	0.0708	0.2449	0.1767	0.4087	0.2429
2007/3/8	0.1198	0.0723	0.2780	0.2034	0.3989	0.2234
2007/3/9	0.1168	0.0642	0.2576	0.1751	0.4022	0.2323
average	0.1054	0.0589	0.2515	0.1794	0.3629	0.2089
primitive=timeframe						
	unexpected	unexpected_r	unexpected	unexpected_r	unexpected	unexpected_r
2007/3/3	0.0327	0.0177	0.0463	0.0301	0.0924	0.0581
2007/3/4	0.0244	0.0123	0.0318	0.0213	0.0629	0.0388
2007/3/5	0.0363	0.0213	0.0543	0.0380	0.1011	0.0646
2007/3/6	0.0339	0.0195	0.0411	0.0272	0.0924	0.0534
2007/3/7	0.0328	0.0219	0.0512	0.0338	0.1033	0.0628
2007/3/8	0.0371	0.0229	0.0418	0.0265	0.1043	0.0582
2007/3/9	0.0303	0.0169	0.0415	0.0271	0.0826	0.0478
average	0.0325	0.0189	0.0440	0.0292	0.0913	0.0548