

組合せ情報フィルタ方式の信頼度を用いた精度改良

有吉 勇介

NEC ヒューマンメディア研究所

情報フィルタリング方式には、情報の内容に基づく CBF (Content Based Filtering) と情報に対する他の利用者の評価に基づく SIF (Social Information Filtering) の 2 種類あり、SIF で選別できない誰も評価していない情報を CBF で選別するという両方式の組合せフィルタ方式も提案されている。しかし、SIF で選別できる情報であっても評価者が少なければ CBF の選別精度の方が良いことが考えられる。そこで本稿では、予測に使用した評価データ数や単語数などから予測評価の信頼度を推定し、それを利用して両方式の組合せを変える方式を提案する。比較実験の結果、従来の組合せフィルタ方式では F 値が 0.33 のところ、提案方式では 0.38 となり、選別精度の改善を確認できた。

Combination information filtering method improvement by using reliability

Yusuke Ariyoshi

Human Media Research Laboratories, NEC corp.

There are two major methods in information filtering area. One is content based filtering (CBF), the other is social information filtering (SIF). CBF can filter information which nobody evaluates. SIF can filter information containing considering figures and tables. The combination information filtering method is based on both CBF and SIF. The author proposes a new combination method that generates higher precision of the final prediction by using the reliability of CBF's and SIF's predictions. New method calculates reliability of prediction from number of learned scores and words, and change combination of CBF and SIF based on reliability. The result of experiments show that F-value is improved from 0.33 to 0.38.

1.はじめに

情報フィルタリングは、情報の洪水の中から利用者が必要とする情報だけを選別する技術であり、利用者に関心に合う新着情報等を推薦する推薦システムの基本技術である [1, 2].

情報フィルタリングは、利用者に関心に合う情報の選別を、利用者が情報に与える評価を予測

することで行う。評価を予測する方式には、キーワードや単語出現頻度などの情報の内容に基づいて予測する CBF (Content Based Filtering) と、情報に対する他の利用者の評価に基づいて予測する SIF (Social Information Filtering) の 2 種類の方式がある。CBF は誰も評価していない情報でも選別可能であり、SIF は図表な

どの価値も織り込んだ選別が可能である。そのため、他の利用者が評価済みの情報はSIFで選別し、SIFで選別できない誰も評価していない情報はCBFで選別する組合せフィルタ方式が提案されている[3]。しかし、SIFで選別できる情報であっても評価者が少なければCBFの選別精度の方が良いことが考えられる。

そこで本稿では、両者を組合せるときに予測評価の信頼度も考慮する方式を提案する。提案方式では両方で評価を予測するとともに、予測に使用した評価データ数や単語数などから予測評価の信頼度を推定し、それを利用して組合せ方を変える。提案方式と従来方式の精度を、技術文書推薦サービスのデータを用いて比較した実験についても述べる。

2.CBFとSIF

情報フィルタリングを使った推薦サービスの流れは次のようになる。まず、システムは、利用者が情報に与える評価を予測することで情報を選別し、利用者に推薦する。そして、推薦した情報に対して利用者の関心に合っているか評価を返してもらい、返ってきた評価から利用者の関心を追加学習することで次のフィルタリングの精度を向上させる。特に、本稿では推薦された情報に対して利用者が段階評価(例えば1~5の5段階)をするタイプのサービスを想定している。

ここでは以下、CBFとSIFという2通りのフィルタリング方式について説明する。

2.1.CBF

CBFは情報内容の特徴をキーワード等で表現したものと、利用者関心をキーワード等で表現したもの(プロフィールと呼ばれる)とを照合することによって情報を選別する。そして、利用者から返ってきた評価をもとにプロフィールを修正することで選別の精度を向上させる。

典型的なCBFとして[4]を例にとると、各情報を出現単語を成分とするベクトル(ドキュメントベクトルと呼ばれる)で表現する。単語Aの成分の値はAが出現していれば1、そうでなければ0とする。プロフィールは、各単語の重要度を成分としたベクトルで表現する。CBFの構成は評価予測とプロフィール学習の2段階から成る(図1)。

評価予測: 現在のプロフィールベクトルを s で、情報 j のドキュメントベクトルを j で表現すると、情報 j に対する予測評価値 \hat{M}_j は、内積 $s \cdot j$ で算出する。

プロフィール学習: 利用者が情報に与えた評価をもとに、その利用者のプロフィールを修正する。プロフィールは利用者関心を単語とその重要度で表現する。プロフィール学習では、利用者の関心に合った情報に含まれる単語の重要度を上げ、関心に合わなかった情報に含まれる単語の重要度を下げる。

例えば[4]では、情報 j に対する実際の評価値が M_j であったとすると、プロフィールベクトル s の更新値 s' を、内積 $s' \cdot j$ が予測値 \hat{M}_j と M_j を $\alpha : (1-\alpha)$ に内分するように決める。ただし、 α は学習レートで α が大ききときは予測評価値 $s' \cdot j$ が実際の値 M_j に近付くように s' が決まり、逆に小さいときは s' が s から離れないように決まる。つまり、学習レート α が小さいほど保守的ということになる。

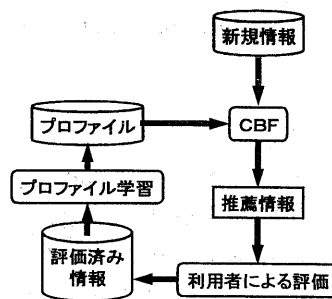


図1: CBFの構成

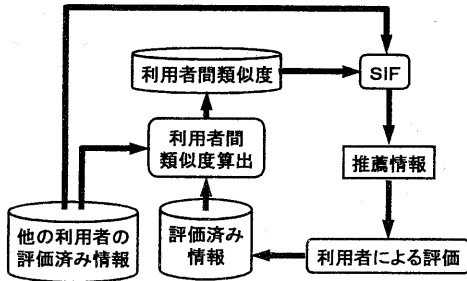


図 2 : SIF の構成

2.2. S I F

S I F のコンセプトは、面白い情報を同じ興味を持つ仲間同士で教え合うという人間が日常行なっている口コミによる情報伝達のシステム化である。典型的な S I F は、利用者が情報に段階評価を与えるタイプで¹、このタイプは利用者間類似度算出と評価予測の 2 段階から成る(図 2)[5]。

$$r_{ii'} = \frac{\sum_{j:i,i' \text{ 共に評価した情報}} (M_{ij} - \overline{M}_i)(M_{i'j} - \overline{M}_{i'})}{\sqrt{\sum_{j:i,i' \text{ 共に評価した情報}} (M_{ij} - \overline{M}_i)^2} \sqrt{\sum_{j:i,i' \text{ 共に評価した情報}} (M_{i'j} - \overline{M}_{i'})^2}}$$

利用者間類似度算出：ここでは、利用者毎に他利用者との類似度を計算する。利用者間類似度は、2 人の評価履歴から共に評価している情報に関する部分を抜き出し、その相関係数を用いる。利用者 i と i' の類似度 $r(i, i')$ は次の式になる。ただし、 M_{ij} は利用者 i が情報 j に与えた評価値であり、 \overline{M}_i は利用者 i が情報に与えた評価の平均値である。

評価予測：情報 j に対する評価の予測は、情報 j に他の利用者が与えた評価値の、利用者間類似度を重みとした重み付平均値によって求める。このとき、類似度が閾値 β 以上の利用者だけを用いることで精度が向上する[5]。利用者 i の情

報 j に対する予測評価値 \hat{M}_{ij} は次の式で求める。

$$\hat{M}_{ij} = \frac{\sum_{i: |r_{ii'}| \geq \beta \text{ である利用者}} |r_{ii'}| M_{i'j}}{\sum_{i: |r_{ii'}| \geq \beta \text{ である利用者}} |r_{ii'}|}$$

3. C B F と S I F の課題と組合せ方式

ここでは、まず、C B F と S I F を次の 2 点、

I) 新規情報に対する選別能力

II) 文書中の図表等などの価値の予測

から比較する。C B F は、i) 誰も評価していない新規情報であっても、その情報が含む単語によって選別することができるが、ii) 図表等の価値は、キーワードや単語頻度からは分からないので織り込むことができない。一方、S I F は、iii) 他人の評価に基づいているため、誰もまだ評価していない新規情報を推薦することはできないが、iv) その情報に価値があるかという人の判断を利用しているため、図表等の価値なども反映されている。

さらに、C B F の長所である新着情報に対するフィルタリング能力について考えると、C B F では、その利用者の評価済み情報から単語の重要度を学習し、その単語重要度を利用して情報を選別している。そのため、v) 同じ新規情報であっても、その利用者の評価済み情報に含まれていない未学習の単語を多く含む情報に対してはフィルタリングの精度は低くなってしまふ。これは目新しい情報ほどフィルタリング精度が低くなるということができ、新着情報推薦を主なサービスとする情報フィルタには望ましくない性質である[3]。

以上の考察から組合せ方式では、予測評価統合とプロファイル学習強化という 2 種類の組合せが提案されている(図 3)。

予測評価統合：C B F と S I F の予測評価を統合する。i) ~ iv) から、基本的には、誰もまだ評価していない新規情報は C B F で選別し、他の利用者が評価済みの情報は S I F で選別するよ

¹ 評価を文章で返すものや、情報を知らせてあげたい利用者のアドレスを返す S I F もある。

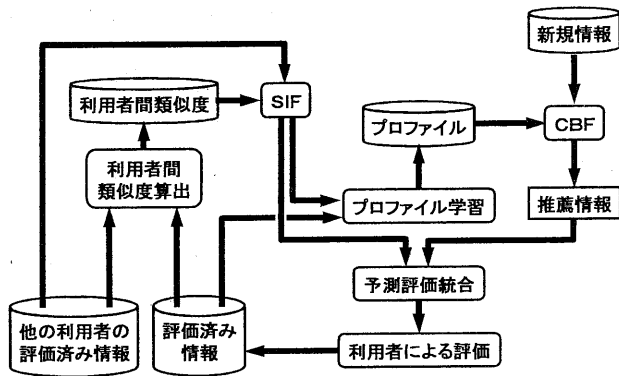


図 3：組合せ方式の構成

うに、予測評価を統合すれば、両方式の欠点を相互補完することができ、それぞれ単独で用いるより選別の精度が向上することが期待できる。プロフィール学習強化：v)の課題を軽減するためにCBFのプロファイル学習でも、SIFと組合せることが考えられる。つまり、プロフィール学習において、その利用者の評価済み情報に加えて、その利用者が未評価情報もSIFによる予測評価を利用して学習する。これにより、その利用者の評価済み情報に含まれないが、SIFで予測評価した未評価情報に含まれる単語の重要度を学習することができるようになり、v)の問題を軽減することができる。

4.提案方式

前述の組合せ方式は、CBFとSIFの欠点を相互補完するように組合せることを提案しているが、具体的なプロフィール学習強化や予測評価統合の方法については述べられていない [3, 6]。提案を素直に実現すると、予測評価統合では、他の利用者が評価済みの情報はSIFで、それ以外はCBFで予測することになる。しかし、その情報を評価した利用者が少人数の場合はCBFのほうが予測評価の信頼度が高いことが考えられる。プロフィール学習強化においても、SIFの予測評価のうち信頼度の高いものだけ

利用したほうが、最終的なフィルタリングの精度が高くなることが考えられる。そこで、提案方式ではSIFとCBFにおいて予測評価に加えてその誤差も推定し、その推定誤差を考慮して組合せる。

4.1.誤差推定

ここではまず、誤差推定の方法を説明する。予測評価は一般に予測に用いたデータの量（例えばCBFではプロフィールの大きさ、SIFでは類似利用者の人数）が多いほど信頼度が高い。そこで提案方式では予測評価の信頼度を測る尺度として予測評価の誤差を用いる。この予測評価の誤差として、本稿では利用者が実際に与えた評価値と予測評価値の差の二乗で計算するものとする。この予測評価の誤差を、提案方式では評価予測に用いたデータの量の多項式で推定する。

4.1.1.CBFの誤差推定

予測評価はドキュメントベクトルとプロフィールから算出されるので、ドキュメントベクトルが大きく、また、プロフィールの学習量が多いほど予測評価の誤差は低いと考えられる。そのため、予測評価の誤差はドキュメントベクトルの大きさと、プロフィール学習量を利用して算出する。

具体的なプロフィール学習量としては、プロフィールに含まれる単語の数を用いる。またドキュメントベクトルの大きさとしては、ドキュメントベクトルとプロフィールの両方に含まれる単語の数を用いる。予測評価の誤差の算出は、ドキュメントベクトルの大きさとプロフィール学習量による次の2次多項式を用いる。

$$\text{推定誤差} = a_1X + a_2X^2 + a_3Y + a_4Y^2 + a_5XY + a_6X' + a_7X'^2 + a_8Y' + a_9Y'^2 + a_{10}X'Y' + a_{11}$$

ただし、Xはドキュメントベクトルの大きさ、Yはプロフィール学習量、X'とY'はそれぞれXとYの逆数である。また、 $a_1 \sim a_{11}$ は多項式の係数である。

実験では、係数 $a_1 \sim a_{11}$ の値は10 Cross Validation (5.2節参照)を行い、実際の誤差にできるだけ近くなるように最小二乗法で決定する。

4.1.2. S I Fの誤差推定

予測評価は利用者間類似度と他の利用者の評価から算出されるので、予測算出に使われた他の利用者が多いほど予測評価の精度は高いと考えられる。そのため、予測評価の誤差は予測評価算出に使われた利用者数、具体的には、利用者間類似度の類似利用者全体での和を使用する。この利用者間類似の和の2次多項式によって予測評価の誤差を算出する。

$$\text{推定誤差} = b_1Z + b_2Z^2 + b_3Z' + b_4Z'^2 + b_5$$

ただし、Zは利用者間類似度で重み付けした類似した利用者の人数、Z'はZの逆数である。また、 $b_1 \sim b_5$ は多項式の係数である。

実験では、係数 $b_1 \sim b_5$ の値は10 Cross Validationを行い、実際の誤差にできるだけ近くなるように最小二乗法で決定する。

4.2.プロフィール学習強化

プロフィール学習強化では、利用者の評価済みコンテンツに加えて、未評価コンテンツからもS I Fによる予測評価を利用して学習する。プロフィール学習には、S I Fによる予測評価のうちから、次の3項目の基準により選択したものを利用する。

1) 実評価と予測評価のバランス：予測評価は誤差を含むので、大量に予測評価から学習する

と誤差の影響が出る恐れがある。そのため、利用者が実際に評価したのと同じ件数だけ、予測評価を利用して学習する。

2) 情報量：利用者の関心を同じ1件の評価から学習するのならば、評価値が中ぐらいのものより極端なものからのほうが良く知ることができる。つまり、予測評価値が大きい方からと小さい方から学習に利用するものを選ぶ。

3) 信頼性：推定誤差が大きいものより小さいものを選ぶ。

上記の基準に基づき、次の2種類の選択方法を考案した。ただしNは利用者が実際に評価した情報の件数である。

1) 単純強化 (バランスと情報量)：予測評価値の大きいものからN/2個、小さいものからN/2個を学習に利用

2) 誤差考慮強化 (単純強化+信頼性)：予測評価値の大きいものからN個、小さいものからN個を選び、その2N個のうち誤差が小さいN個を学習に利用

4.3.予測評価統合

最終的な予測評価を、S I FとC B Fそれぞれによる予測評価から、推定誤差を考慮して決める。具体的には次の2つの方式が考えられる。

1) 切替え統合方式：誤差の小さい方を最終的な予測評価にする。(S I F予測評価の推定誤差/C B F予測評価の推定誤差)が閾値 γ より小さい場合はS I Fによる予測評価を、大きい場合はC B Fによる予測評価を最終的な予測評価に選ぶ。

2) ブレンド統合方式：C B FとS I Fによる予測評価を誤差の比で内分する値を最終的な予測評価にする。つまり、 $(S I Fによる予測評価 \times C B F予測評価の推定誤差 + C B Fによる予測評価 \times S I F予測評価の推定誤差) / (S I F予測評価の推定誤差 + C B F予測評価の推定誤差)$ によって最終的な予測評価を算出する。

5.比較実験

5.1.実験データ

データ収集：評価データを収集するために、社内にて技術文書の推薦サービスを行った。推薦サービスでは推薦文書に対し利用者に関心に応じて1(なし)から5(あり)の5段階評価をしてもらった。

実験データ：実験では、収集したデータのある程度継続利用した45人の利用者のもに絞った。実験に使用したデータは、文書数は3000文書で、評価データは10709件、利用者当たりの平均評価数は238.0件/人、文書あたりは平均3.57件/文書となり、利用者と文書の全組合せのうち利用者が実際に評価したものは7.93%になる。評価値の分布を表1に示す。

ドキュメントベクトルの生成：ドキュメントベクトルは、各文書の題名と要約から名詞の可能性のある語のみを抽出した。また、数詞等は除いた。ドキュメントベクトル中の単語数は表2の通りである。

5.2.実験方法

実験方法：実験では10 Cross Validationによって評価指標の算出やパラメータの決定を行った。

10 Cross Validationは評価データを10ブロックに分割し、1つのブロックを残り9ブロックのデータから予測することを、全10ブロックについて行う。その結果、利用者の実評価データ全件に対して予測評価値が算出されるので、そこから評価予測能力の分析が行える。

後述のパラメータの決定では指標としてF値を使用した。F値は適合率と再現率の2つの指標を

統合したもので、 $2 \times \text{適合率} \times \text{再現率} / (\text{適合率} + \text{再現率})$ で定義される。ここで、適合率と再現率は、利用者が実際に高い評価を与えた情報を正解情報、システムによる予測評価値が高い情報を推薦情報としたとき、

適合率 = 推薦情報全体中の正解情報の割合

再現率 = 正解情報全体中の推薦情報の割合

と定義する。「高い評価」の基準によって、

厳しいサービス：システムは四捨五入した予測評価が5の情報に推薦し、利用者が評価5を与えたものを正解とする。

緩やかなサービス：システムは四捨五入した予測評価が4以上の情報を推薦し、利用者が4以上の評価を与えたものを正解とする。

と想定するサービスが異なるが、技術文書推薦サービスの利用者どちらのサービスを望むかインタビューしたところ、希望はどちらかに偏らず、同じ利用者でも状況や気分によって希望は変わることが分かった。そのため、基準を5とした場合のF値と4以上とした場合のF値の平均を指標にした。

パラメータ：CBFとSIFのアルゴリズムは2節で述べた従来方式を使用した。そのため3種類のパラメータ、CBFの学習速度 α 、SIFの閾値 β 、切替え統合の閾値 γ 、を決める必要がある。そこで、パラメータを変化させながら10 Cross Validationで評価予測を行うことで、パラメータを決定した。その決定手順は、フィルタリング処理の流れにそって、SIFの閾値 β 、SIFの誤差推定係数を決め、学習強化方式毎にCBFの α 、CBFの誤差推定係数、切替え統合の γ という順番とした。 α 、 β 、 γ は10 Cross Validationでの評価予測のF値が最大になると

表2：評価分布

評価数	評価点数分布				
	1	2	3	4	5
10709	5165	1997	1547	1270	730

表1：ドキュメントベクトルの単語数

文書数	単語数		
	平均	最大	最小
3000	32.58	66	10

きの値を採用し、誤差推定係数は 10 Cross Validation での評価予測の推定誤差が最良になるように最小二乗法で決定した。パラメータ決定と評価は別々のデータを用いることが望ましいが、今回は利用者が少ないため同じデータを用いた。

決定したパラメータ値とそのときのF値が表3である。ただし、S I F / 従来統合の列のF値はS I F で予測できない評価値のみC B F を利用した場合のF値である。

5.3.実験結果

表3からF値を見ると従来の組合せフィルタ方式では単純強化・従来統合で約0.33のところ、提案方式では誤差考慮強化・切替え統合で約0.38となり、フィルタリング精度の改善が確認できる。

図4～7は10 Cross Validation によって予測した評価値を、評価値の大きい順に取ったときの適合率とその時の順位を表している。情報フィルタリングはもっぱら推薦サービスに使うため、予測評価値の大きい部分での適合率が重要である。表1から今回のデータでは実際の評価値が5のものは730件、4以上のものは2000件あるので、順位が700～2000辺りの適合率が重要になる。

統合をしないC B Fのみで見ると誤差考慮強化

が一番良く(図4)、ブレンド統合方式では単純強化が(図5)、切替え統合では誤差考慮強化が一番良くなっている(図6)。どの場合でも誤差を考慮して統合もしくは学習強化した方式が最良となっており、誤差を考慮したほうが従来より適合率が良いことが分かる。また、表3からはプロファイル学習強化では誤差考慮したほうが従来よりF値が良いことが分かる。

各図で最良の方式を比較すると(図7)、誤差考慮強化では切替え統合とC B Fのみは、ほとんど同じになっている。これは、この切替え統合ではS I Fの推定誤差が0になる3件以外はC B Fが選ばれているのためである。切替え統合の誤差考慮強化とブレンド統合の単純強化を比べると、順位が上段のグラフでは約1300、下段では約1100より高いときは僅かに切替え統合の誤差考慮強化が良く、それより低いときは僅かにブレンド統合の単純強化が良くなっており、両者の得意な範囲が異なることが分かる。

6.おわりに

C B FとS I Fを組合せた組合せフィルタ方式の改良として、S I FとC B Fにおいて予測評価に加えてその誤差も推定し、その推定誤差を考慮して組合せる方式を提案した。また、実験により従来方式より適合率が高いことを確かめた。

S I Fはある程度のフィルタリング精度を出すためには利用者が最低数百人必要だと報告されている[5]。今回の実験では利用者は45人と少なくS I F単独の性能はC B Fに比べ低いものであったが、誤差を考慮してC B Fと組合せることで高い性能が出る事が分かった。これにより、従来はS I Fで情報推薦サービスをすることができなかった人数の少ない集団に対しても、本稿の方式を用いればサービスすることができると考えられる。

なお、本研究は日本情報処理開発協会の次世代

表3：パラメータ

上：パラメータ 下：F 値	C B F のみ	S I F/ 従来統合	切替え 統合
	α	β	γ
強化無し	175.0	0.46	0.6896
	0.3291	0.2763	0.3374
単純強化	123.2	0.46	0.7941
	0.3062	0.3282	0.3076
誤差考慮強化	108.6	0.46	0.0
	0.3751	0.3282	0.3752

電子図書館システム研究開発事業の一部として行われている。

参考文献

- [1] 森田, 速水, “情報フィルタリングシステム”, 情報処理 37(8), pp.751-758, 1996.
- [2] “Special Section: Recommender Systems”, CACM, Mar 97, Vol.40, No.3, pp.56-89, 1997.
- [3] 有吉, 市山, “情報の内容と他者の評価を利用した情報フィルタリング方式”, 電子情報通信学会第8回データ工学ワークショップ論文集, pp.49-54, 1997.

- [4] Kivinen, M. Warmuth, “Exponentiated Gradient Versus Gradient Descent for Linear Predictors”, *J. UCSC-CRL-94-16*, 1994.
- [5] Upendra Shardanand, Pattice Maes, “Social Information Filtering: Algorithms for Automating “Word of Mouth””, Proc. of CHI'95, pp.210-217, 1995.
- [6] Marko Balabanovic, Yoav Shoham, “Fab: Content Based, Collaborative Recommendation”, *Communications of the ACM*, Vol.40 No.3, pp.66-72, Mar 1997.

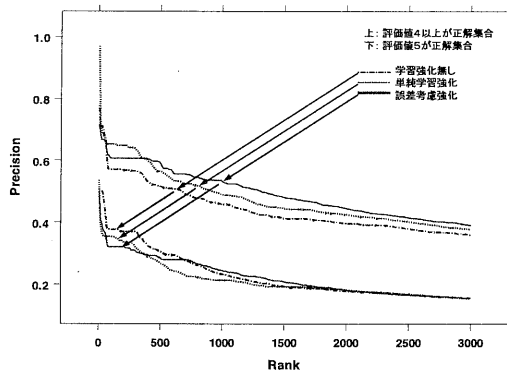


図 4: 適合率-順位: CBFのみ

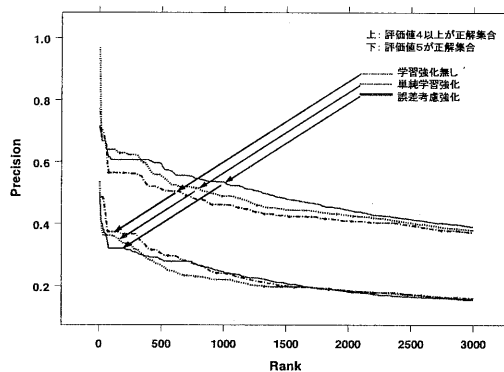


図 6: 適合率-順位: 切替え統合

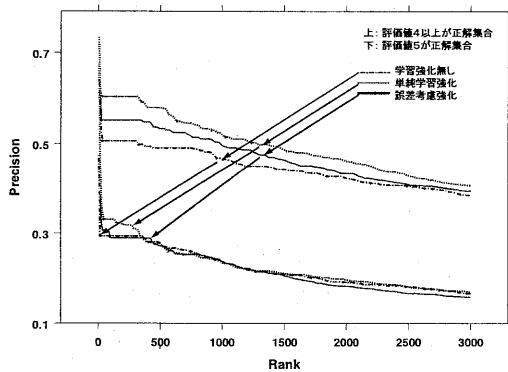


図 5: 適合率-順位: ブレンド統合

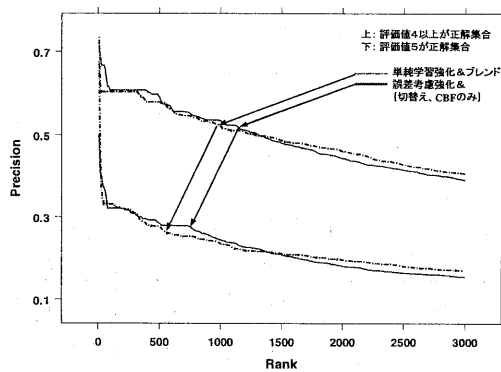


図 7: 適合率-順位: 図1~3での最良方式