閲覧者の目的と属性に応じた技術資料の推薦

田口　浩† 坂上　聡子† 岩田　雅史†

† 三菱電機株式会社 先端技術総合研究所　〒 661–8661 兵庫県尼崎市塚口本町 8–1–1
E-mail: † {Taguchi.Hiroshi @ dw, Sakajo.Satoko @ ds, Iwata.Masafumi @ ak}.MitsubishiElectric.co.jp

あらまし　本論文では，蓄積された資料の中から必要な資料を検索するユーザに対し，目的に合った資料を予測して推薦する推薦型技術資料検索システムを提案する．本提案は，製品の保守担当部門など，新旧の技術資料が大量に蓄積されている現場への適用を対象とする．提案するシステムでは，他のユーザの閲覧履歴を，技術資料閲覧時の特徴を踏まえて分析することにより，閲覧者の目的や属性によって異なる閲覧傾向をとらえる．そのうえで，ショッピングサイトなどにおいて実用化実績のある協調フィルタリングを応用した予測と推薦を行う．これにより，ユーザに新たな負担をかけることなく，適切な資料の推薦を可能とする．

キーワード　情報推薦，協調フィルタリング，保守作業支援

Personalized Document Recommendation for Field Engineers

Hiroshi TAGUCHI†, Satoko SAKAJO†, and Masafumi IWATA†

† Advanced Technology R&D Center, Mitsubishi Electric Corporation
Tsukaguchihonmachi 8–1–1, Amagasaki-shi, Hyogo, 661–8661 Japan
E-mail: † {Taguchi.Hiroshi @ dw, Sakajo.Satoko @ ds, Iwata.Masafumi @ ak}.MitsubishiElectric.co.jp

Abstract　In this paper, we propose a personalized document recommendation system. The system aims to support field engineers maintain products which are operated over several decades. A maintenance section has a large number of old and new documents for maintenance. The system recommends suitable documents to each engineer according to its needs. The method features grasping tendencies of reading by reader's purpose and skills from reading history. It gives a recommendation based on the tendencies by collaborative filtering. The results of applying the method to actual data have shown it can give suitable recommendation.

Key words　recommendation, collaborative filtering, maintenance work

1. はじめに

電気設備や冷凍空調設備など，設置されてから数十年にわたり稼働を続ける製品は，定期点検や故障対応といった保守作業が不可欠である．それらの作業に備え，新機種の発売や作業方法の改善の度に多数の技術資料が作成されており，保守技術者はそれらを適宜活用する．しかし，膨大に蓄積された資料の中から目的に合う資料をすぐに探し出すことは容易ではない．

本論文では，資料を検索するユーザに対し，目的に合った資料を予測して推薦する推薦型技術資料検索システムを提案する．提案に先立ち，我々は実データを用いて技術資料閲覧時の特徴について考察した．本システムでは，考察で得られた知見を踏まえて他のユーザの閲覧履歴を分析することで，閲覧者の目的と属性によって異なる閲覧傾向をとらえる．そのうえで，協調フィルタリングを応用した予測と推薦を行う．これにより，ユーザに負担をかけず，必要に応じた資料を推薦できる．

2. 保守現場における技術資料管理

2.1 技術資料検索システム

保守担当部門では，図 1 に示すような，技術資料を電子ファイルとして管理する技術資料検索システムを有することが多い．保守技術者は Web を通じて，資料が格納されているサーバにアクセスし，必要な資料を検索して閲覧する．Web サーバはアクセスの履歴を示すログを記録して保存する．

技術者が資料を必要とする際に，実施する頻度が高い定型的な作業に関する資料であれば，求める内容が明確であり，目的の資料を比較的容易に探し出せる．しかし，故障修理などの非定型作業に関する資料を必要とする場合は，技術者自身が求める内容を明確に表現できないことも多い．さらに，一見関連が浅そうな資料にヒントとなる記述が含まれていることも珍しくない．そのため，特に初めて経験する故障修理などにおいて，適した資料を短時間で探し出すことは困難である．

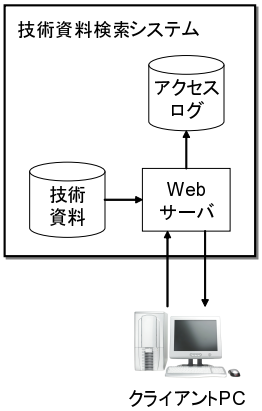

図 1　技術資料検索システム
Fig. 1　Document retrieval system.

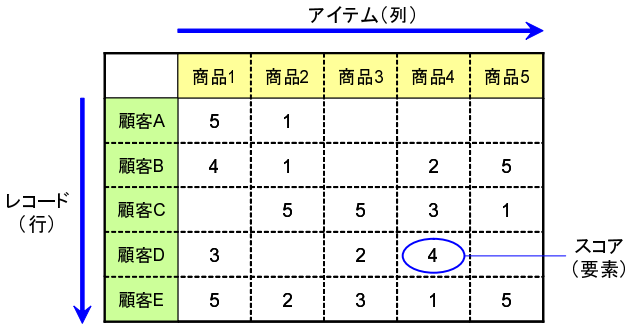

図 2　評価値行列
Fig. 2　Rating matrix.

このような作業と資料との対応付けは，経験によって知ることができる．しかし，そのノウハウをルール化することや，大量の資料をそのような観点で系統立てて管理することは現実的ではない．そのようなことを行わずとも，技術者が必要とする資料を予測し，検索を支援する手法が求められる．

上で示した技術資料検索システムのように，資料が電子ファイルとして管理されている場合，それらのファイルへのアクセスログを得ることは容易である．アクセスログからは個人ごとの閲覧履歴が分かり，資料閲覧の傾向をとらえることができる．その情報を活用して場合に応じた資料を予測できる仕組みがあれば，検索にかかる時間を短縮させられ，作業の効率化を支援できる．

2.2　協調フィルタリング

ユーザの今後の行動を，過去の行動や嗜好が似ている他のユーザの情報をもとに予測する手法として，協調フィルタリングがある．協調フィルタリングは，電子商取引システムなどにおいて多く実用化されており，代表的な例として Amazon 社のショッピングサイトにおける商品推薦機能 [1] がある．これらのシステムでは，協調フィルタリングを用いて，顧客が各々の未購入商品を今後購入する可能性を，その顧客と他の顧客の購入履歴や商品評価履歴から求め，購入可能性が高い商品を推薦する．この例に照らし合わせると，協調フィルタリングの特長として以下の点があげられる．

- 顧客による初期入力が不要．
- 運営者による商品群のコンテンツ情報の管理が不要．
- 他の情報推薦技術に比べ，思いがけない推薦が可能．

一般的に，協調フィルタリングによる予測を行う際は，入力として用いる行動履歴情報を，図 2 に示すような評価値行列の形式で表現する．行は行動の主語，列は推薦対象物に相当する．要素はその行動の特徴を表す数値であり，未知の場合は空欄とする．例えば，商品評価履歴を行動履歴情報として用いるとすれば，行は顧客，列は商品，要素は評価値となる．嗜好情報などの行動履歴以外の情報も，この形式に変換可能であれば入力に加えられる．本論文では，評価値行列の行をレコード，列をアイテム，要素をスコアと呼ぶ．協調フィルタリングでは，評価値行列において空欄となっているスコアを予測する．

協調フィルタリングを実現する方式は複数提案されている．代表的なものとして，ユーザベース方式，アイテムベース方式，Slope-One アルゴリズム方式がある．

ユーザベース方式 [2] は，予測対象レコードと他の各レコードとの類似度を求め，類似度が高いレコードで高スコアのアイテムほど，予測対象レコードでも高スコアと予測される．レコード間の類似度は，各アイテムのスコアが似ているほど高くなる．

アイテムベース方式 [3] は，アイテム間の関連度を求め，予測対象レコードで高スコアのアイテムと関連度が高いアイテムほど高スコアと予測する．関連度は，両アイテムともものスコアが既知であるレコード数などに基づいて求める．つまり，先に述べた商品推薦機能の例で説明すると，ある 2 つの商品を両方とも購入している顧客が多いほどその商品間の関連度は高いという考え方による予測方式である．

Slope-One アルゴリズム方式 [4] は，各アイテム間のスコアの差は全レコードで同じであると仮定してスコアを予測する．全レコードから各アイテム間のスコアの平均偏差を求めておき，予測対象レコードで既知のスコアとその平均偏差から，未知のスコアを予測する．

3.　資料推薦場面の分析

3.1　商品推薦と資料推薦の相違点

協調フィルタリングを用いることで，これまでの閲覧履歴において，現在と似た状況のときにどのような資料が役立っていたかという情報に基づき，現在の状況に応じた資料の推薦が可能になると期待される．ここで，協調フィルタリングを技術資料の推薦に適用するには，一般的に適用されている電子商取引での場面とはいくつか異なる点があることを考慮しなければならない．以下に，本事例において着眼すべき事項をあげる．

現在の担当業務および知識や経験の考慮が必要　保守業務の内容は多岐にわたり，各技術者が毎回同じ業務を担当するとは限らない．それゆえ，技術者の過去の行動履歴や特徴だけを考慮すればよいわけではなく，現在の担当業務に応じた技術資料を推薦する必要がある．また，技術者ごとに知識や経験の程度が大きく異なるので，この点も考慮しなければならない．商品推薦では，過去の購入履歴や評価履歴のみに基づいて推薦することが一般的である．

技術者ごとの閲覧傾向は担当してきた業務によって変わる　協

調フィルタリングでは，入力情報で既知のスコアはその時点での確定値として扱われる．商品推薦において商品評価履歴を入力に用いるとすると，顧客の商品に対する評価値であるスコアはすぐに大きく変わる可能性は少ないので，確定値と見なすことができる．しかし，技術者の資料に対する閲覧回数や閲覧時間といった値は確定値とはいい難い．例えば，閲覧回数が少ないからといって，その資料がその技術者にとって役に立たなかったというわけではなく，その資料を必要とする業務を担当する機会が少なかったためとも考えられる．

過去に閲覧したことがある資料も推薦候補となりうる　商品推薦では，顧客が過去に購入経験のある商品を推薦することの効果はあまり大きくないので，未購入の商品の中から推薦商品を決めることが多い．これに対し，技術資料推薦では，かなり前に担当した業務を久しぶりに担当している場合など，過去に閲覧したことのある資料の推薦が有効な場合も少なくない．

3.2　閲覧目的の区切りと資料の有益度

閲覧履歴に基づいて資料の推薦を行うためには，その閲覧履歴から，どういった状況のときに各資料がどの程度役立ったかという情報を整理しなければならない．どの程度役立ったかを表す度合いを有益度と呼ぶこととする．上で述べた点を踏まえると，同じ資料であっても閲覧の目的によってその有益度は異なるので，同一目的による閲覧ごとに各資料の有益度を求める必要がある．そこで，閲覧目的の区切りと資料の有益度について，実データに基づく分析を行った．

ある作業のために資料を閲覧する際に，同時または連続して閲覧する資料は，ともにその作業と関連し，その資料どうしにも何らかの関連がある可能性が高い．このように連続して資料を閲覧している間は，同一目的による閲覧と考えることができる．それゆえ，閲覧目的の区切りは，ある資料を表示してから次の資料を表示するまでの時間で判断できると考えられる．

この点を分析するため，ユーザ数が 1,000，資料数が 2,036 の 1 日分の閲覧履歴を用いて，同一の閲覧者がある資料を表示してから次の資料を表示するまでの時間の分布を調べた．その結果，10 分未満が全体の 93.8%，10 分以上 30 分未満が 4.4%，30 分以上 1 時間未満が 0.6%であった．10 分未満の場合が大半を占めるが，10 分以上 30 分未満の場合も 30 分以上の場合に比べるとその割合は多い．これより，資料を熟読する場合は 10 分以上 30 分未満の時間をかけると推測される．よって，前の資料を表示してから 30 分未満で次の資料を表示している場合は同一目的での閲覧が続いていると見なし，30 分以上経過していれば新しい目的による閲覧に移行したと見なすことができる．

次に，資料の有益度について分析するため，ユーザ数が 1,685，資料数が 2,054 の 1ヶ月分の閲覧履歴を用いて，同一の目的内での閲覧回数と閲覧時間の分布を調べた．ここで，閲覧目的の区切りとなる時間は，上述の分析結果に従って 30 分とした．

閲覧回数はアクセスログから正確な値を得られるが，図 3 に示すように，1 回が全体の 73.5%，2 回が 22.3%で，平均は 1.35 回となり，偏りが大きい．そのため，資料間の差を表しにくい．一方，閲覧時間は，図 4 に示すように，1 分未満が全体の 33.4%，1 分以上 2 分未満が 13.6%で，平均は 6 分 9 秒となり，閲覧回数と比較して偏りが小さい．閲覧時間にはユーザが真に見ていない時間が含まれる場合もあるが，有益度を求める指標として閲覧回数よりも妥当であると考えられる．

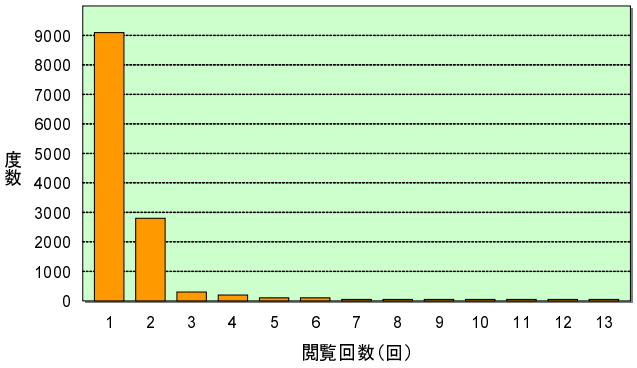

図 3　閲覧回数の分布
Fig. 3　Distribution of reading frequency.

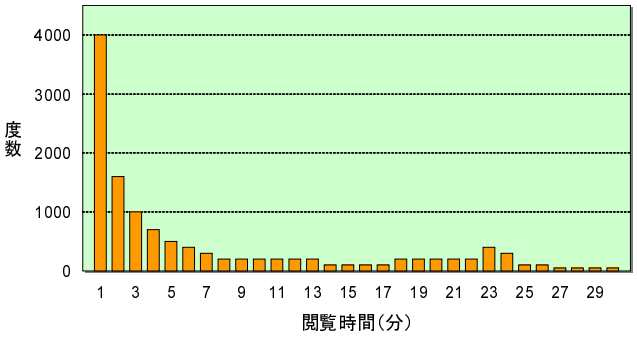

図 4　閲覧時間の分布
Fig. 4　Distribution of reading duration.

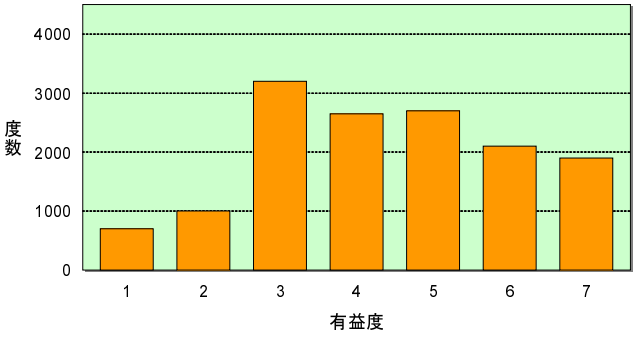

図 5　閲覧時間に応じた有益度の分布
Fig. 5　Distribution of ratings based on reading duration.

閲覧時間に基づいて有益度を求めるうえで，ある情報に対してユーザの興味の度合いが高いほど閲覧時間も長くなるという相関関係を利用する．例えば，文献 [5] では，Web ページの閲覧時間とユーザの関心度の関係を分析した結果,「大変興味があった」と評価したページの平均閲覧時間は 40.7 秒,「少し興味があった」は 14.9 秒,「興味がなかった」は 5.4 秒であったと報告されている．この報告において，関心度の高い順に有益度を 3，2，1 としたとき，有益度を x，閲覧時間を y 秒として近似曲線を求めると，$y = 1.97\mathrm{e}^{x}$ で表される曲線となる．これを踏まえ，閲覧時間 t 秒に応じて，以下の数式を満たす s を有益度とする正規化を試みる．

$$\begin{cases} s = 1 & (0 < t < 2) \\ 2 \times \mathrm{e}^{s-1} \leqq t < 2 \times \mathrm{e}^{s} & (2 \leqq t < 1800) \end{cases} \quad (1)$$

これにより，1〜7 の 7 段階の有益度に正規化され，その分布を図 5 に示す．図 4 と比較して正規分布に近い分布となっており，各資料の有益度を適当に表すことができているといえる．

3.3 技術者属性

技術資料の推薦においては，さらに技術者の知識や経験の程度も考慮する必要がある．それらを技術者属性と呼び，その一例を以下にあげる．

技術評価 社内の統一基準による技術者全体での総合的な位置付けを等級で示す．

所有資格 特定の業務を担当するために必要な社内資格のうち，取得済みのものの一覧を示す．

担当分野 業務を「電気設備」,「冷凍空調設備」などの大きな分野に分類したときに，担当している分野を示す．

保守担当部門では，個々の技術者が担える業務を明らかにするうえで技術者属性は重要な情報である．そのため，システムで管理されている場合が多いが，技術資料を管理するシステムとは別に構成されていることが一般的である．このようなシステムを人材情報管理システムと呼ぶこととする．人材情報管理システムを技術資料検索システムと連携させることにより，技術者属性を考慮した資料推薦が実現可能になると考えられる．

4. 推薦型技術資料検索システム

4.1 目的ごとの閲覧傾向を示す評価値行列の作成

3 章で述べた分析に基づき，検索者が必要とする資料の予測と推薦を行う推薦型技術資料検索システムを提案する．図 6 にその構成を示す．本システムは，従来の技術資料検索システムに，協調フィルタリングによる推薦を行う推薦エンジンを追加した構成となっている．推薦エンジンでは，アクセスログに基づく閲覧履歴と，人材情報管理システムから得られる技術者属性情報を連携させることで，閲覧者の目的と属性によって異なる閲覧傾向をとらえたうえで推薦を行う．

推薦エンジンではまず，閲覧履歴を分析し，協調フィルタリングの入力となる評価値行列を作成する．一般的な協調フィルタリングでは，評価値行列の 1 レコードは 1 ユーザの履歴情報を表す．しかし，保守技術者は多様な作業を行っており，履歴情報を技術者ごとに整理すると，作業内容に応じた閲覧傾向を埋没させてしまう．それゆえ，閲覧履歴を閲覧者ごとではなく閲覧目的ごとに区切り，それをレコードの単位とする．図 7 に示すように，ある資料を表示してから次の資料を表示するまでの時間が 30 分以上の箇所を閲覧目的の区切りとする．そして，閲覧目的の区切りごとに，その間の履歴情報を 1 つのレコードに整理する．評価値行列のアイテムは推薦対象となる各資料であり，資料に対するスコアは各レコードにおける有益度とする．3.2 節で示した正規化方法を適用し，閲覧時間に基づいて 1〜7 の 7 段階で有益度を求める．

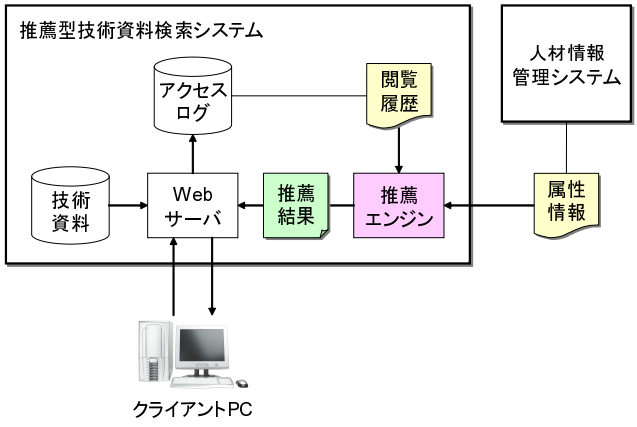

図 6 推薦型技術資料検索システム
Fig. 6 Document recommendation system.

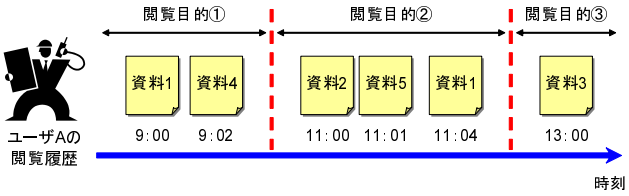

図 7 閲覧目的の区切り
Fig. 7 Separation between reading purposes.

これにより，目的ごとの閲覧傾向を反映した予測が行えるとともに，推薦対象者が過去に閲覧したことのある資料も推薦候補に含めることができる．

4.2 閲覧者の属性を反映する評価値行列の拡張

続いて，閲覧履歴における各閲覧者の技術者属性を反映するために評価値行列の拡張を行う．この拡張では，人材情報管理システムで管理されている情報に基づき，アイテムとして技術者属性を追加する．図 8 に示すように，技術者属性をアイテムに含めることで，知識や経験の程度も考慮した推薦が行える．

協調フィルタリングで技術者属性情報を扱うためには，それらを数値で表現し，かつ，その数値が間隔尺度，すなわち，数値の大小およびその差の比が意味を持たなければならない．したがって，そのように表現されていない情報を用いるには数値化が必要である．数値化には，以下の方法が考えられる．

- 「技術評価」といった，順序尺度などで表現されている情報は，間隔尺度となるように正規化する．
- 「所有資格」といった，項目が列挙されている情報は，その項目数に変換する．

	資料アイテム					技術者属性アイテム		
	資料1	資料2	資料3	資料4	資料5	年齢	技術評価	所有資格数
閲覧1	2			4		37	14	9
閲覧2	1	1			3	37	14	9
閲覧3			5			37	14	9
閲覧4		3	1			28	9	4
閲覧5	4			1	2	42	15	7

閲覧目的ごとに区切って整理したレコード

図 8 技術者属性を追加した評価値行列
Fig. 8 Extended rating matrix.

- 「担当分野」といった，属性値が択一式の情報は，個々の属性値をアイテムとして分割し，該当する属性値のスコアを 1，該当しない属性値のスコアを 0 とする．

4.3 推薦の実行

ここまでで作成される評価値行列を入力として協調フィルタリングによる予測と推薦を行う．レコードを同一目的による閲覧ごとに整理しているため，これから資料を検索するユーザに推薦を行うためには，その閲覧を表すレコードを 1 行加える必要がある．追加レコードでは資料アイテムのスコアがすべて空欄となるので，最初に閲覧する資料については推薦を行わず，ユーザ自身が任意で選ぶこととする．そして，2 番目の資料を検索する際に，最初の資料のスコアを更新して協調フィルタリングを行うことにより，他の資料のスコアを予測し，予測値が高い資料をユーザに推薦する．

このように本システムでは，膨大な資料に新たなタグ付けを行ったり，ユーザに事前入力を求めたりすることなく，ユーザの閲覧目的と知識や経験の程度を考慮した資料推薦が実現される．

5. 実データへの適用

提案システムでの資料推薦手法を実データに適用し，その結果を考察した．用いたデータは，3.2 節の分析でも用いた 1ヶ月分の閲覧履歴である．

5.1 評価値行列のスパース性

まず，作成された評価値行列について考察する．作成された評価値行列は，レコード数が 4,508，アイテム数が 2,061 で，空欄でないスコア数は 11,035 となった．アイテムのうち，資料アイテムは 2,054，技術者属性アイテムは 7 である．今回は技術者属性として，年齢，技術評価，所有資格，担当分野，技術普及活動回数の情報を用いた．このときの，空欄のスコアの割合を示すスパース率は 99.9%である．

協調フィルタリングを用いる際には，スパース性，すなわち，評価値行列における未知スコアの多さが問題となる．スパース率が大きいと予測の性能が下がるとされている．協調フィルタリングの性能評価に度々用いられる MovieLens データセット [6] のスパース率は，レコード数が 6,040，アイテム数が 3,900 の評価値行列において 95.8%であり，適切なスパース率の目安のひとつになるといえる．

この目安と比較すると，作成された評価値行列のスパース率は大きすぎる．そこで，スパース率を低減させるように，評価値行列を修正した．まず，ほとんど閲覧されない資料は，過去の閲覧履歴において有益であった可能性が低いと考えられるので，その資料をアイテムから除く．除外する資料の基準を，スコアが既知のレコード数が 20 未満の資料とすると，除外後の評価値行列は，レコード数が 1,997，アイテム数が 85，空欄でないスコア数は 3,448 となり，スパース率は 98.0%となった．さらに，閲覧した資料の少ないレコードは予測にあまり貢献できないと考えられるので，そのようなレコードも除外する．除外するレコードの基準を，スコアが既知の資料アイテム数が 4 未満であるレコードとすると，除外後の評価値行列は，レコード数が 88，アイテム数が 72，空欄でないスコア数が 423 となり，スパース率は上述の目安を満たす 93.3%となった．

5.2 推薦結果

次に，修正された評価値行列を用いて推薦を行った結果について考察する．ここでは，最も多くのレコードでスコアが既知である資料 A を最初に閲覧した場合の，次に閲覧すべき資料として上位 3 つの資料を推薦することとした．2.2 節で述べた，協調フィルタリングの異なる 3 つの実現方式をそれぞれ用いて推薦を行ったところ，以下の結果が得られた．

- アイテムベース方式では，資料 A の対象機種と非常に似た機種についての類似資料が推薦された．非常に関連が高い資料が推薦されたといえる．
- ユーザベース方式では，資料 A の対象機種と同系統の機種についての類似資料が推薦された．アイテムベース方式ほどではないが，比較的関連がある資料が推薦されたといえる．
- Slope-One アルゴリズム方式では，資料 A と内容が大きく異なる，関連が浅そうな資料が推薦された．

協調フィルタリングの実現方式が異なると推薦結果も異なることから，適切な評価値行列が構成できているといえる．推薦されたそれぞれの資料が有益であるかは，ユーザの求める内容に大きく依存するので客観的に判断することは難しい．しかし，最初に閲覧した資料から推測される閲覧目的に合致すると思われる資料も推薦されていることから，本手法は適切な資料推薦を実現できるといえる．

6. おわりに

本論文では，製品の保守担当部門などを対象とした，推薦型技術資料検索システムを提案した．提案するシステムでは，他のユーザの閲覧履歴から閲覧者の目的と属性を考慮した閲覧傾向をとらえたうえで，協調フィルタリングの応用による予測と推薦を行う．本推薦手法を実データに適用した結果，適当な推薦結果が得られることを確認した．

今後は，提案したシステムの実現場における有効性の検証を行う予定である．

文　献

[1] G. Linden, B. Smith, and J. York, “Amazon.com Recommendations: Item-to-Item Collaborative Filtering,” IEEE Internet Computing, vol.7, no.1, pp.76–80, Jan. 2003.

[2] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, “GroupLens: An Open Architecture for Collaborative Filtering of Netnews,” Proc. 1994 ACM Conference on Computer Supported Cooperative Work, pp.175–186, Oct. 1994.

[3] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Item-Based Collaborative Filtering Recommendation Algorithms,” Proc. 10th International World Wide Web Conference, pp.285-295, May 2001.

[4] D. Lemire and A. Maclachlan, “Slope One Predictors for Online Rating-Based Collaborative Filtering,” Proc. SIAM International Conference on Data Mining, April 2005.

[5] 清水勇喜，“WWW におけるユーザの興味対象と閲覧時間の関係の調査，” 情報処理学会第 57 回全国大会講演論文集（3），pp.191–192，Oct. 1998.

[6] GroupLens, “MovieLens Data Sets,” http://www.grouplens.org/node/73, Oct. 2006.