

# 第 4 章

# 4

## 業務の流れを効率化する



## 第4章 業務の流れを効率化する 目次

	情報研究所	主任研究員	高橋 光裕
	情報研究所	主任研究員	大屋 隆生
	情報研究所	主任研究員	二方 厚志
本部	研究企画グループ	主任研究員	桑畑 暁生
	情報研究所	主任研究員	堤 富士雄
	情報研究所	主任研究員	三浦 輝久
	情報研究所	上席研究員	篠原 靖志
	情報研究所	主任研究員	田中 真人

4 - 1	情報システム技術	43
4 - 2	知識マネージメント技術	51
コラム4	電力会社とe-ビジネス	50
コラム5	大量文書自動整理システム「トピックうおっちゃん」	58



高橋 光裕（1985年入所）  
入所以来、ソフトウェア開発・保守の見積り技術、コスト・導入効果の定量的評価技術、システム監査・開発管理技術、研究に従事してきた。  
現在は、組織としてのIT活用の成熟度を診断する技術を研究している。



堤 富士雄（1994年入所）  
オフィス文書、写真、点検映像、設備音など、さまざまな情報を扱う検索システムのユーザインタフェースに関する研究に従事してきた。現在は、大量のデータを効率的に閲覧できるインタフェースの研究に取り組んでいる。



大屋 隆生（1984年入所）  
発電限界費用計算手法、経営情報システム、電磁界の健康影響に関する疫学調査の動向分析、システム導入効果評価、技術選択手法、IT活用地域ビジネスなど、ORや統計分析による情報数理に関する研究に従事。



三浦 輝久（2001年入所）  
入所以来、データマイニング手法の開発に従事。効果的なデータの分析を可能とするデータベースを構築する手法として、的確な予備の調査と調査結果の詳細な分析に基づくデータベース構築法の確立を目指している。



二方 厚志（1987年入所）  
入所以来、協同作業支援、情報共有、情報の組織化等の研究に従事。現在、顧客ニーズの獲得に関する研究に取り組んでいる。



篠原 靖志（1984年入所）  
入所以来、専門家の知識の自動抽出、文書の検索支援・自動整理など、機械学習技術に基づく情報収集・分析手法の研究開発に従事。現在、データマイニングのためのデータベース構築法、不確実性を考慮した最適運用のための確率予測手法などの研究に取り組んでいる。

桑畑 暁生（12ページに掲載）



田中 真人（株 電力計算センターより出向）  
EWSやPC向けにGUIアプリケーション開発、比較的小規模なデータベースを利用するソフトウェアの開発及びWeb関連でCGIスクリプト等の作成などに従事している。

## 4 - 1 情報システム技術

### 4-1-1 情報システム開発規模評価

#### 背景・目的

情報システムを新規開発・保守するに当たっては、開発規模を適切に評価（見積り）することが非常に重要である。開発規模は、開発プロジェクトに必要な資源（人・物・金・時間）を割り当てるコスト見積りの最も重要な根拠である。したがって、開発規模の評価が過少であったり、仕様変更などによって開発規模が大きく増えるなどすると、開発に必要な資源が不足し、プロジェクトに混乱が生じたり、納期遅れ・品質の低下などを引き起こす原因となる。

情報システムの規模を測る尺度としては、記述されるプログラムの行数（SLOC：Source Lines of Code）が伝統的に使われてきた。しかし、1990年代にクライアント/サーバ・システムやWebアプリケーションなど、開発技術の多様化が進んだ結果、1システムで複数のプログラミング言語を混用したり、プログラミングなしで使えるツールやパッケージが増えるなど、SLOCを簡単に計測したり集計することが次第に困難になってきた。また、利用する開発技術が変わると規模の値が大きく変わる、実装方法が決まらなると計測（見積り）できないという点でも、SLOCは情報システムの大きさを表すには不適切となった。

これらSLOCの問題を解決する尺度として国内外で近年急速に普及しつつあるのが「機能規模」尺度である。

#### (1) ファンクションポイント法と機能規模

ファンクションポイント法（FP法、Function Point Analysis）は、1970年代末に米国IBM社のA. J. Albrecht氏が提案した機能規模計測手法である。

FP法で計測される機能規模は、システムの具備する機能の量を利用者の視点から計測した値であり、具体的には、情報システムが、利用者にとって論理的に何種類の情報を入出力したり管理したりするかに着目し、その種類数を点数化するものである。ここで、論理的とは、**図4-1-1**に示すように、画面・紙面の制約、性能向上な

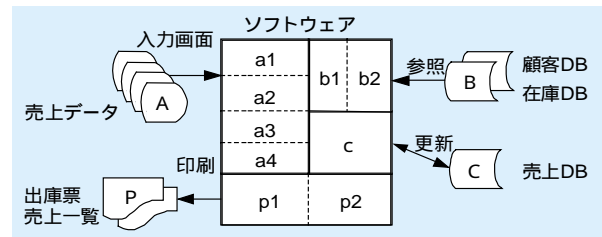


図4-1-1 FP法の概念

どの実装上の都合で分割/統合される前の、利用者にとって意味のある情報の塊を指す。

機能規模計測では、システムの物理構成（モジュール構成、ファイル構成など）、システム内部での処理ロジック（処理手順や計算ロジック）、品質要件などは一切考慮されない。このため、機能規模には、

機能要件が同じならば、品質要件や技術要件が違ってても値が変動しない、

推定するのではなく規則に従って計測をするため、誰が何回計っても同じ値が算出できる、

利用者の視点で計測するので利用者でも計測結果を理解することが容易、

という長所がある。

計測された機能規模は、開発プロジェクトの所要工数や品質データ（欠陥発生数、所要テスト件数）などとの相関が高いことが世界中での利用経験から知られている。

機能規模の具体的な計測方法としては、電中研法（日本）、IFPUG法（米国）、MK-II法（英国）、NESMA法（オランダ）、COSMIC-FFP法（カナダ他）など、さまざまな計測手法が各国で提案され、普及している。

これらの手法は、ISO/IECによって定められた機能規模計測に関する国際標準（ISO/IEC 14143 - 1）に適合した手法であるが、各手法による計測結果はそれぞれ異なった値になり、手法間での互換性はない。

#### (2) 当所におけるファンクションポイント法への取り組み

##### a. 電中研法FP

当所では、1986年からFP法の研究に着手し、日本の情報システム開発の特徴を考慮した計測手法として、電

中研法を開発した。電中研法（第2版）の計測手順を図4-1-2に示す。

この電中研法は、以下の特長を持つ。

入出力媒体の違いを考慮し、表4-1-1に示すファイル、電文、画面、帳票の媒体ごとに異なる基準値を与えている。

計測労力を軽減するために、項目数などを厳密に数えずに複雑度を判定する簡素化した方法を用いている。

計測洩れや誤りを防ぐために、機能を洩れなく洗い出すために、機能階層分割法という独自の機能列挙手法を提供している。

計測範囲を決める際に、ソフトウェア階層を考慮し、アプリケーション層とシステムソフト層を分けて計測する。

組込/制御系のソフトウェアを計測するための拡張規則を備えている。

システム保守における保守規模を計測するための拡張規則を備えている。

システム特性（品質要件など）の違いを加味する補正を廃止している。

このうち、電中研法FPにおいて採った品質要件の評価をFP法から切り離す方針（上記）は、その後、ISO/IECにおけるFP法の標準化でも採り入れられている。

当所は、電力会社などの依頼を受け、電中研法を特定企業向けにカスタマイズすることも行っており、電中研法およびそのカスタマイズ版は、電力会社などで利用されている。

なお、当所におけるFP法研究の成果は、FP法ハン

表4-1-1 電中研法（第2版）の基準値

機能要素の型		複雑度別FP基準値		
媒体	形態	単純	普通	複雑
ファイル	入力	3点	5点	8点
	出力	3点	5点	8点
	入出力	3点	6点	13点
電文	入力	3点	5点	8点
	出力	3点	5点	8点
	制御	2点	3点	5点
画面	入力	3点	6点	13点
	出力	3点	5点	8点
	入出力	6点	11点	19点
	メニュー	2点	3点	5点
帳票		3点	6点	10点

ドブックとしてとりまとめられており、この中ではFP法の導入・運用指針や利用ガイドなども示されている。

#### b. 開発規模算定のための品質要件の加味

機能規模は、品質要件と技術要件を考慮しない尺度であるため、所要工数を見積りする際には、品質要件や開発要件の違いを加味した開発規模を算出する必要がある。

そこで当所は、独自に開発した品質チェックリストを用いて品質要件を評価し、機能規模を補正して「価値量（価値を考慮した規模）」を算出し、それを経て開発量を算出する「段階的規模評価体系SMISS（Stepwise Measurement of Information System Size）」を開発した。

このSMISSでは、情報システムの品質を、主特性6、副特性19、補助特性38の3階層の特性体系に分類し、この体系に沿って作られたチェックリストを用いて情報システムの品質水準を評価することとしている。

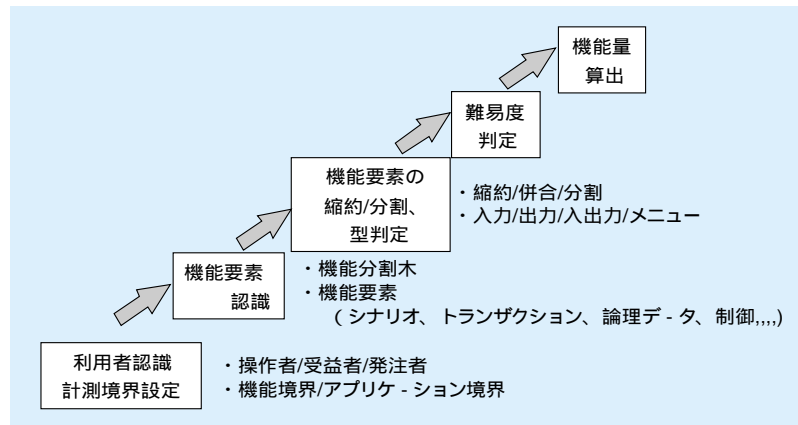


図4-1-2 電中研法FPの計測手順

## 4-1-2 情報システム導入効果評価

### 背景・目的

電気事業では1990年代に、オフィス生産性向上や組織活性化を目指してパソコンやLANなどの社内情報化インフラの導入を進めてきたが、その目的は従業員の積極的な利用により達成される。そこで、新商品への消費者の反応を分析し、販売を拡大する方策を考えるマーケティング手法を応用し、社内情報インフラに対する従業員の意識の調査分析にもとづいてインフラ導入効果を評価する手法と、その利用促進方策を探索する図4-1-3に示す評価モデルを開発した。

### アンケート調査による評価

ある電力会社では業務改善（業務の効率化、ペーパーレス化、情報の活用、コミュニケーション、高度情報化社会への対応）をねらいとして、パソコンとLANの全社的な展開を行った。平成8年1月から本格導入を開始し、平成8年3月末までに、全体の約半数（本店・支店の全部と一部の営業所）に1人1台のパソコンを導入した。本格導入開始直後（約1ヶ月）と約半年後に利用に関するアンケート調査を行い、試験導入参加の効果、アプリケーション利用状況とその効果などを評価した。

#### a. アプリケーション利用度分析

本分析では、利用決定要因として、取り決めの有無（部または課での利用の取り決め事項を認識しているか

どうか）試験導入参加（試験導入時から参加しているかどうか）年齢（40歳未満かどうか）の3つを取り上げた。

従業員の各アプリケーションの「利用」・「未利用」の判断は、電子メール、スケジュール、共用情報DBにより、また電子掲示板については利用実績により行った。電子メールとスケジュールについては、月に5回以上、共用情報DBと電子掲示板については月1回以上の利用を「利用」とした。

表4-1-2に導入直後における各アプリケーションの利用者率に対する、各利用決定要因のオッズ比を示す。オッズ比とは、その要因のある従業員の間での利用者率を $p_1$ 、その要因のない従業員の間での利用者率を $p_0$ とすると、それぞれのオッズ（非利用者に対する利用者の割合） $p_1/(1-p_1)$ と $p_0/(1-p_0)$ の比 $[p_1/(1-p_1)]/[p_0/(1-p_0)]$ である。

表4-1-2で、\*が付いたオッズ比にかかわる要因は、95%の統計的有意性を持って、そのアプリケーションの利用促進要因であり、#が付いたオッズ比にかかわる要因は95%の統計的有意性を持って、そのアプリケーションの利用阻害要因であることを示す。取り決めは、全般に利用促進に有効な方策と考えられる。共用情報DBと電子掲示板において、試験導入参加者が非参加者に比べ、利用が少ないのは、試験導入時と本格導入時で、使用するソフトウェアを入れ替えたことが原因と考えられる。電子メール、電子掲示板については、年齢の高い（40歳以上の）利用者の利用率が低くなっており、これらを重点とした高年齢層への利用支援方策が必要であった。

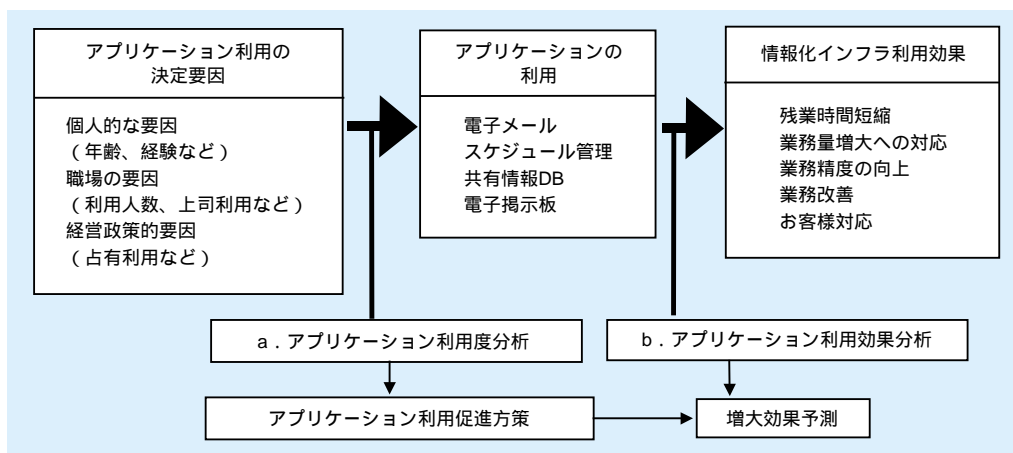


図4-1-3 情報化インフラ効果評価モデル



表4-1-2 利用決定要因による利用者率のオッズ比（本格導入直後）

アプリ名 利用決定要因	電子メール	スケジュール	共用情報DB	電子掲示板
取り決め有り	1.36 (2.09 - 0.88)	3.11* (5.14 - 1.89)	2.16* (3.83 - 1.21)	1.44 (2.55 - 0.81)
試験導入参加	1.71* (2.29 - 1.28)	2.09* (2.85 - 1.54)	0.33 (0.48 - 0.23)	0.44 (0.63 - 0.30)
年齢 (40歳未満)	2.67* (3.51 - 2.04)	0.60 (0.79 - 0.46)	0.88 (1.17 - 0.66)	1.44* (1.96 - 1.06)

数字は、上段がオッズ比、下段が95%信頼区間（上限 - 下限）

表4-1-3に本格導入半年後の各アプリケーションの利用者率に対する、各利用決定要因のオッズ比を示す。表4-1-2より、取り決めの有無は、スケジュール管理と共用情報データベースの利用に大きな効果があることがわかる。どのアプリケーションでも試験導入参加者は先進的に利用している。特に導入直後の利用の決定要因を分析した表4-1-2では、ソフトウェアの入れ換えが原因となり共用情報データベースと電子掲示板で試験導入参加者の利用者率が低くなっていたが、約半年後の利用者率は高くなっており、ソフトウェアの入れ換えにも適応したと考えられる。高年齢（40歳以上）の従業員は、若い（40歳未満）従業員と比較して、電子メールでは利用者率が低くなっているが、他のアプリケーションでは有意な差はみられなくなっている。

#### b. アプリケーション利用効果分析

2回のアンケートでは各アプリケーション毎に、「そのアプリケーションの利用による生産性の向上が時間に換算すると、1日あたり何分になるか（省力化時間）」を設問し、省力化時間が、利用頻度が増加するに従って有意に増加しているかどうかを検定した結果、以下が明

らかになった。

2回のアンケートのどちらにおいても、どのソフトでも利用頻度が高いほど省力化時間の平均値が長くなっており、アプリケーションの利用により効果（省力化時間）が実感できていることを示している。

本格導入の約半年後には、直後と比較して、従業員が感じる情報化インフラによる業務効率化（省力化時間）が大きくなっている。これは、

- ・各アプリケーションの利用率が高くなったこと、
- ・同じ利用頻度でも、利用者が感じる業務効率化の程度が大きくなったこと、

の相乗効果によるものであると確認できた。

#### 4-1-3 業務支援用情報システム

##### 背景・目的

企業を取り巻く社会環境の変化にともない、企業内の個人が担当する業務内容は広範囲かつ複雑なものへと変わりつつある。このような状況下における業務では、従来までの情報の内容による整理・検索だけでなく、業務を行なう上で情報がどのような組み合わせで利用される

表4-1-3 利用決定要因による利用者率のオッズ比（本格導入半年後）

アプリ名 利用決定要因	電子メール	スケジュール	共用情報DB	電子掲示板
取り決め有り	0.82 (0.32 - 2.07)	7.15* (3.08 - 16.58)	2.25* (1.17 - 4.34)	1.41 (0.64 - 3.12)
試験導入参加	2.97* (1.79 - 4.95)	3.62* (2.06 - 6.38)	2.47* (1.51 - 4.04)	2.64* (1.59 - 4.38)
年齢 (40歳未満)	2.50* (1.52 - 4.09)	1.49 (0.91 - 2.41)	1.39 (0.89 - 2.16)	1.06 (0.68 - 1.67)

数字は、上段がオッズ比、下段が95%信頼区間（上限 - 下限）

のかを把握し、業務に即した形で情報提供する仕組みが重要になってくる。そこで、本節では、情報の組み合わせと提供について当所で行なってきた研究の概略を紹介する。

### (1) 作業状況に基づくファイル間のリンクの自動生成

協調作業から生まれたさまざまな文書を関連付けて整理するために、ファイル間の関係をタイプ付きリンクとして表現し、ファイルとは独立に一元管理する仕組みを開発した。これを「自己増殖型データベース」と呼んでおり、図4-1-4にイメージを示す。自己増殖型データベースでは、リンクのタイプを、仕事の流れ、分類、作成順、詳細化、参照による連携、の5種類に分類している。自己増殖型データベースは、

- ・リンクタイプによるファイル間の関係の容易な把握、
  - ・アプリケーションと連携したリンクの自動生成、
  - ・ファイル削除等に対するリンクの自動修復、
- といった特長を持つ。

たとえば、図4-1-4のように自己増殖型データベースと研究進捗管理システムを連携させた場合、ワークフローに対応するリンク（仕事の流れ）が自動作成される以外にも、ワークフロー内の共同作業で作成された詳細資料Fや、インターネットから検索した文献H、G等を

リンクで結び付けることができ、過去の作業内容に基づいた情報の発見・再利用を行なうことが可能になる。

### (2) 作業状況に基づくファイル群の自律的な整理

特定の作業でよく一緒に使われるファイルを、階層的なグループとして自動的にかつ継続的に整理する自己組織化機構を開発した。本機構を図4-1-5に示す。本機構では、あるファイルが一定時間内に他のファイルと一緒に操作された回数から、ファイル間のリンクとその強度を計算する（図4-1-5、 $G_1$ 、 $G_2$ 、 $G_3$ ）。そして、一緒に操作されることが多いファイルを自動的にグループ化する（図4-1-5  $G_1$ 、 $G_2$ 、 $G_3$ ）。個々のグループは、たとえば「計画立案段階の初期に一緒に利用された」といった、業務中の個々の作業で用いられるファイル群の関係を表わしている。さらに、リンク強度を集約した値を使ってグループ自らが階層的にグループを形成していくことで、たとえば、「初期の立案段階で使われたファイル群  $G_1$  とその後の作業で使われたファイル群  $G_2$ 」といった、業務での作業の流れに対応する関係を表現する（図4-1-5  $M_1$ 、 $M_2$ ）。本機構の特徴は以下の通りである。

- ・ある計画の立案段階の資料といった過去の作業と結びついた形での情報の再利用が可能になる。
- ・組織改変や作業手順の変更にグループ構成が適応して

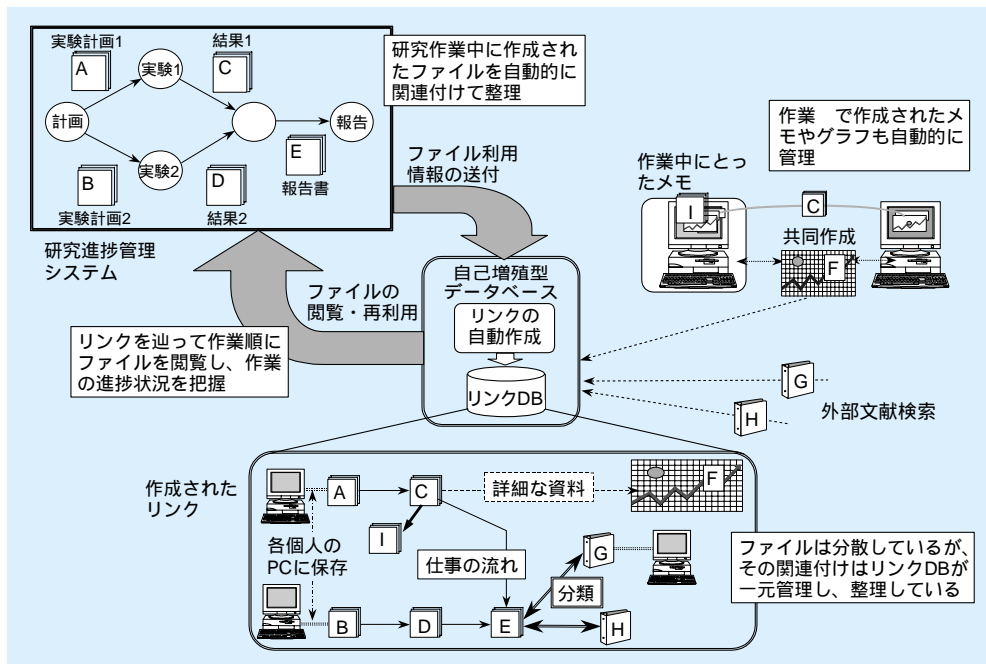


図4-1-4 自己増殖型データベースのイメージ

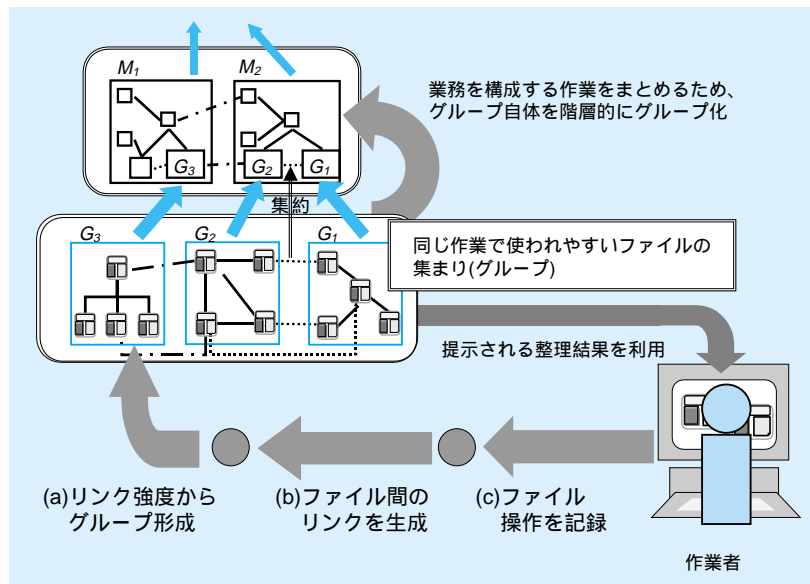


図4-1-5 ファイル群の自己組織化過程

いくため、現状に即した情報活用が可能になる。

研究作業を題材とした実験を行ない、作業中に作成された2つの文書と参考にした資料から、作成文書を中心とした2つのグループが適切に形成されることを確認した。また、シミュレーション実験の結果からは、1つの作業に行なうファイル操作回数が3回以上であれば適切なグループが形成されるであろうとの結論を得た。

### (3) 作業状況に応じた利用ファイルの予測

ユーザが利用中のファイル群から次に必要となるファイル群を予測するためのルール（予測ルール）を過去のファイル利用記録から抽出する手法を開発した。本手法は、たとえば、平成14年の「案件1」の計画立案中に「計画書を利用したとき、メモも利用する」という傾向を、

（H14、計画、案件1）（H14、メモ、案件1）

といった予測ルールとして抽出し、「計画書を利用している場合には次にメモが必要になる」といった予測に用いる。本手法では、ファイルを（H14、計画、案件1）のような属性値の3つ組（アイテム）として一般化して表現する。また、「H14」や「案件1」、「計画」といった作業状況を表わす属性値による分類（コンテキスト）毎にルールの抽出を行なうことで、作業によって異なるファイル利用パターンをルールに反映させる。本手法の特長は以下の通りである。

- ・ 個々のファイルではなく、ファイルの属性に基づく一般性があるルールを抽出するため、ルールの適用可能

性を高めることができる。

- ・ コンテキストの導入により、作業状況に応じた重要なルールを抽出できる。
- ・ 文献（電中研研究報告：R02011）の手法と組み合わせることで、利用傾向が似ている利用者のファイル利用ルールを用いた予測ができる。

### (4) ルール抽出実験

実作業からのルール抽出実験を行い、アイテム全体からルール抽出を行なう通常手法（コンテキスト無し）との比較から、以下の結果を得た。

#### 予測性能の向上:

コンテキスト無しの通常手法と比べ、総合的な予測性能（F値）<sup>(注1)</sup>が向上することを確認した。結果を表4-1-4に示す。

#### 確信度<sup>(注2)</sup>による絞り込み:

予測されたアイテムから確信度が高い上位 $n$ 個を選択

表4-1-4 提案手法による予測性能の向上

選択方式	再現率	精度	F 値
通常手法	0.71	0.54	0.61
提案手法	0.95	0.52	0.68

(注1) 再現率：予測されるべきアイテムが予測結果中に現われる割合。精度：予測結果中のアイテムが予測されるべきものである割合。F値は再現率と精度の調和平均であり、 $F 値 = 2 \cdot 再現率 \cdot 精度 / (再現率 + 精度)$ 。

(注2) 確信度：予測が正しい確率。



した場合、予測アイテム中の 不要アイテムと 選択されなかった必要アイテム（取りこぼしアイテム）を考慮すると、図4-1-6に示すように、 $n = 3$ から4の間でバランスの良い情報提供が可能になることを確認した。

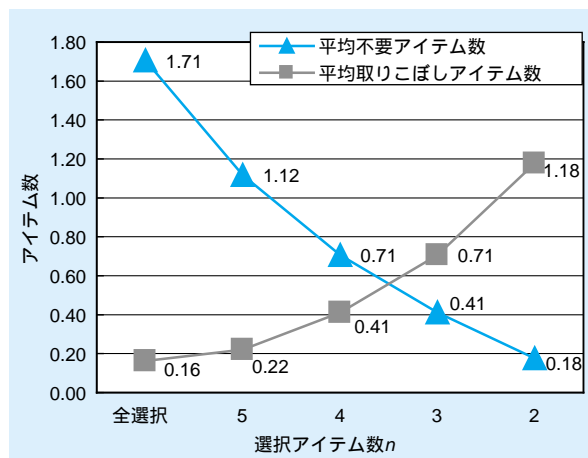


図4-1-6 上位候補選択の効果

## コラム4：電力会社とeビジネス

eビジネスの概念はそれ自体、新しいこともあって、多様な意味をもつが、広義の意味ではインターネット、電子データ交換（EDI：Electric Data Exchange）、CALs（Commerce at Light Speed）等を利用した電子商取引、情報提供、広告、マーケティングなどの活動全般と捉えられる。そのeビジネスのエネルギー産業での代表的事例（ここでは米国が中心）を見ると、目的別に資材調達、電力取引、顧客サービスの3つに大別できる。

### 資材調達

資材調達（SCM：Supply Chain Management）に関する代表例はeマーケット・プレイス（e-Market place:EMP）である。EMPは商品を販売するチャネルが弱い小規模事業者等と、良質の資材を安価に調達したい企業とを結びつける市場である。特定分野（化学、医薬、自動車など）ごとに特化したEMPはパーティカル型、幅広い分野の商品を取り扱うEMPは水平型と称されるが、米国電力会社21社が中心となりコンソーシアムを形成して立ち上げたパンテロス(Pantellos)は不振といわれるBtoB市場において成功したパーティカル型EMPの一つとして評価されている。

### 電力取引市場

電力取引市場は市場自由化により出現する一種

のEMPである。先行する米国では大手電力会社の発電、配電部門は、ほぼ全て電力取引市場に参加している。各社電力取引部門は系統連係のない他地域での取引（2003年3月時点では全米に公営・市営を含めて13ほどの市場が存在）にも積極的に参加しており、そのため現在、各電力会社は市場ごとに異なる取引ルールに対応した各種取引システム、社内の財務管理・決済システムを用意した上で各市場に参加している。

### 顧客サービス

顧客サービスに向けた例としては、Webでのサービス申込み・オンライン支払・エネルギー情報（履歴、利用状況）の提供、省エネコンサルティング、コールセンタ等が挙げられる。エネルギー情報の提供については顧客がオンラインで登録することで、これまでの電力使用量や支払いに関する履歴を閲覧できるサービスを提供する社が多く、その中でもビジネスユーザ向けに簡単なエネルギー利用解析、比較などを提供するエネルギー・マネジメント・システム（EMS）を提供する会社が目立つ。また、コールセンタによる苦情処理、事故処理についても地理情報システム（GIS）とコールセンタ、社内業務管理システムを連携させ、作業進捗状況の情報提供ができるようなシステム構築例が見られる。

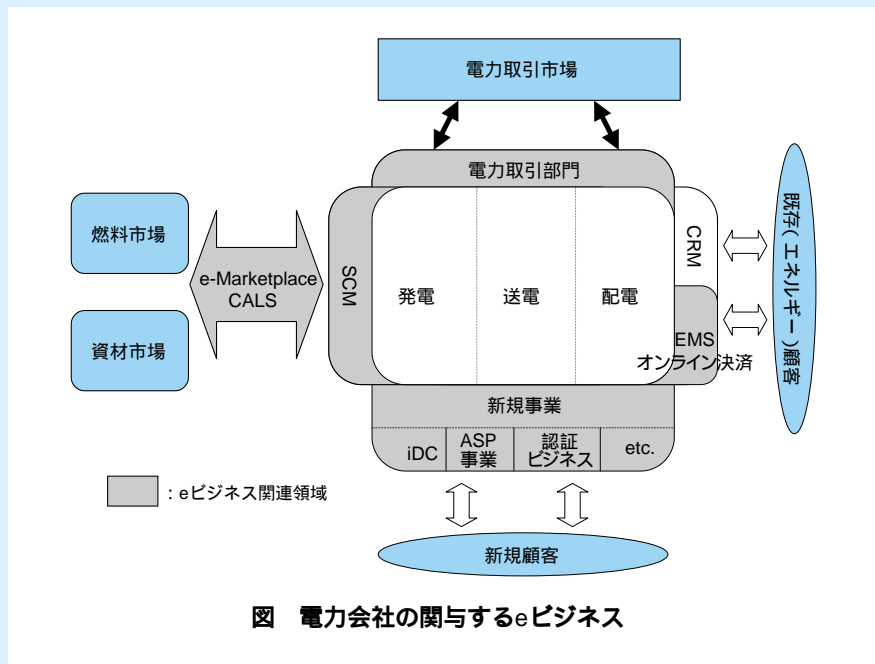


図 電力会社の関与するeビジネス

## 4-2 知識マネージメント技術

### 4-2-1 文書・画像データベース

#### 背景・目的

電子技術の発展により、未整理かつ多量の文章・画像などの電子情報がオフィスに溢れている。これらの情報を活用し業務高度化に結び付けるには、情報検索やデータベースの構築において、ユーザ側の負担をできる限り少なくするユーザフレンドリーなインターフェースが必要である。

#### (1) ユーザフレンドリーな文書検索システム

未整理の文書情報を扱う代表的な方法として google (<http://www.google.com>) に代表される全文検索がある。しかし、従来の全文検索では、検索条件として論理式表現が採用されており、その使いこなしの難しさが指摘されていた。また表記揺れや利用者の曖昧な記憶により、部分的に間違ったキーワードが使われた場合の検索精度低下が課題であった。そこで、この問題を解決する新たな検索方式を開発した。これはキーワード検索につづり曖昧検索を加えた新しい検索手法であり、図4-2-1に示すように、キーワードを画面中にマウスで配置し、



図4-2-1 キーワードを2次元配置する文書検索システム

その配置を検索条件として用いるという特徴的なユーザインターフェースを備えている。

図4-2-1の画面下部が検索条件を指定する部分で上部に検索結果が表示される。画面上のキーワード(画面下部右上付近の「マルチメディア」「電力設備」)はマウスによって自由な位置に移動できる。キーワード位置の上下方向がそのキーワードがどの程度重要か(優先度)を表し、左右方向がつづりの曖昧さをどの程度許容するかを表す。例えば上にあるキーワードは下にあるキーワードよりも、検索結果として得たい文書に含まれていて欲しい度合いが強いことを表す。また右にあるキーワードは正確で、つづりの曖昧さを考慮する必要がないことを表し、左にあればつづりが少々違っていても許容することを表す。

提案方式の有効性評価のために、5万件の文書群と11人の被験者を対象とした比較実験を行った。その結果、代表的な従来手法である論理式を用いる方式(論理式型)および短い文章を検索入力とする方式(文書入力型)よりも、多くの関連文書を安定に検索できることがわかった。

図4-2-2に実験結果の一例を示す。これは行われた全検索7792回を検索方式別に重み付き得点率と呼ぶ評価基準で整理した結果である。重み付き得点率とは、最も適切な検索結果を100%とした場合の相対的な結果の良さを表し、例えば0%とは検索結果の上位に関連文書が1件もないことを表す。

図から従来の論理式型や文書入力型では、全検索に占める低い重み付き得点率の割合が大きく、90%以上の良い検索結果が少ないことがわかる。これに対し、提案方式は低い重み付き得点率の結果が少ない上に、高い重み付き得点率の検索結果が多く、当たり外れが小さく安定して良い結果を出せる検索方式であることがわかる。

#### (2) 設備巡視点検用映像データベース

電力設備の保守・管理の合理化を目指し、従来の定期保守・点検から機器状態に依存した保守・点検への移行が検討されている。この実現には十分な量の整理された設備状態情報が必要となる。この情報取得の一手段とし

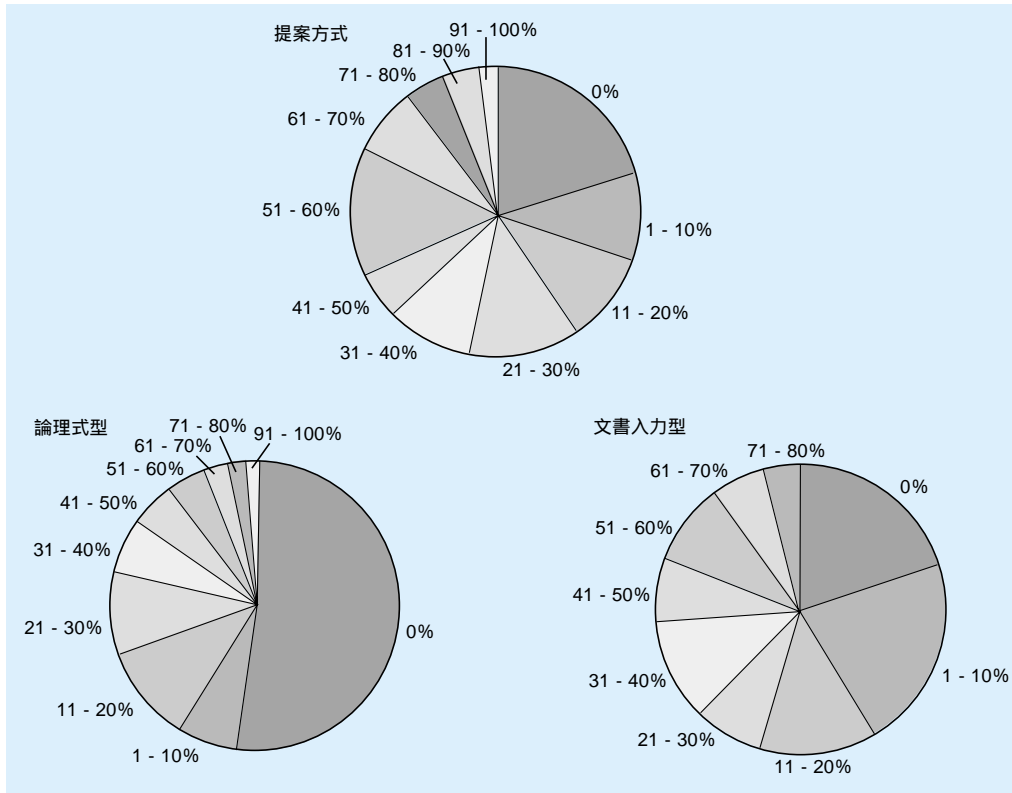


図4-2-2 全検索の重み付き得点率での分布

て巡視点検作業の活用が考えられるが、運転中の多様かつ多量な情報を収集するには多くの手間とコストを要するため、効率的な収集技術が必要である。ところが、作業者が機器状態の判断に活用している五感情報を、作業者に大きな負担をかけることなく低コストで収集し活用する技術はこれまで存在しなかった。そこで、巡視点検時の映像情報を、効率的に収集・分類・蓄積すると共に、その場で過去の類似例（機器状態）を検索・表示することで巡視点検作業を支援する「設備巡視点検用データベース」の実現を目指して、その構築のキーとなる以下の手法を開発した。

#### a . 未編集映像のシーン分割手法

機器の点検映像を連続記録した未編集のビデオ映像を、点検対象機器が映っているシーンや作業を行っているシーンなど、映像内容に対応した複数のシーンに分割する手法を開発した。シーンは、映像変化に応じて動的に閾値を変動させる手法をベースに分割され、パラメータ制御により、適切なシーン分割を行う。

屋外の電力施設を撮影した映像、および実験室内での作業映像を用いて従来手法との比較を行い、機器や作業

手順を正確に分割できるなど、一般用途のビデオ編集で使用されている従来型シーン分割では不可能な分割を行えることを示した。

図4-2-3に設備の分解点検を模擬し、小型装置の分解映像を提案手法でシーン分割した結果をアルバム状に表示している。作業時間の長短や、作業動作の細かさの違いにかかわらず、ネジの取り外しや、部分装置の取り外しなど分解手順を正確に反映した結果となっている。

#### b . 迅速な登録と高精度な認識を両立させるビデオ映像認識手法

機器の持つ色情報をベースに、設備点検などの現場で、利用者がビデオ映像を容易に登録したり、点検対象機器と類似あるいは同一の機器を容易に検索できる方式を開発した。

開発方式の有効性を確認するために、色情報を使う代表的なビデオ映像認識手法と比較評価した結果、色情報を扱う手法として最も優れた手法であることを明らかにした。さらにプロトタイプシステムを開発し、提案方式が迅速に機能することを確認した。



図4-2-3 装置分解映像のシーン分割結果

## 4-2-2 データベース拡充手法

### 背景・目的

激しい競争を勝ち抜くために、企業内に大量に蓄積されたデータから、企業の様々な意思決定を助ける知識を抽出するデータマイニングの必要性は非常に高まってきている。データマイニングにより、役に立つ知識を抽出するには、抽出の元となるデータベースの整備が重要である。データベースが不十分な場合、有用な知識を抽出することは不可能であり、質のいいデータベースの構築が、効果的なデータマイニングの鍵となる。

しかしながら、例えば、顧客情報データベースを作ろうとする場合、それぞれのお客様の住所、年齢、家族構成、電気使用料等、さまざまな収集可能なお客様の属性すべてを含んだデータベースを構築することは、データ収集に予算制約がある場合、非常に難しいと思われる。そこで、既存のデータベースを基に、予算を考慮しつつ、有用な知識が得られるように、データベースを徐々に拡充する手法を開発した。

### (1) クリティカルデータ特定法

予算を考慮しつつデータベースの整備を行う方法とし

て、まず既存のデータから関連しそうな情報を抜き出した初期の顧客情報データベースを準備し、必要に応じ新たな属性データを外部から収集し、効果的な知識が抽出できるようにデータベースを徐々に拡充することが考えられる。ここで問題となるのは、追加すべき属性をどのように決定すればいいかということである。データ追加の際のデータ取得コストと、データ追加から得られる利益の間の費用対効果を考えると、単純に最も良い知識が得られる属性を追加すればいいわけではなく、その属性データの入手に必要なコストを考慮して追加属性を選択しなければならないということがわかる。

当所では、この課題を解決するために、クリティカル・データ特定法を開発した。開発手法の特長は、追加すべき属性を決定するために、効果的な予備調査を設計し、実施するという点である。分析者にとって、外部データは大体どのような特性のデータであるのかは推定可能であっても、実際に入手するまで、データの内容を知ることにはできない。分析者の推定が間違っていた場合、コストを払い入手した外部データが結果的に知識の改善につながらないといった危険性が生じる。開発手法は、少数のお客様に対する予備調査を実施することで、すべてのお客様のデータを見ることなく、精度の高い知識の抽出に役立つ外部データを特定し、無駄なデータ取得コ



ストを省く。予備調査では分析システムと分析者が協調して、追加属性を特定する。

開発手法の流れを図4-2-4に沿って説明する。まずシステムは、現在の顧客情報データベースを解析することで、不足しているデータの情報を分析者に提示する。一方、分析者は、システムが提示した情報をもとに、予備調査を実施する(同図A)。次に、予備調査で入手した情報をもとに、システムは各属性追加後の抽出知識の精度を推定する。分析者は推定精度とデータ収集コストの費用対効果の評価を行い、追加属性を決定する(同図B)。最後に、分析者は本格調査により追加属性のデータを収集し、顧客情報データベースを拡充する(同図C)。図4-2-5のように、予備調査と本格調査を繰り返すことによって、データベースを拡充していく。

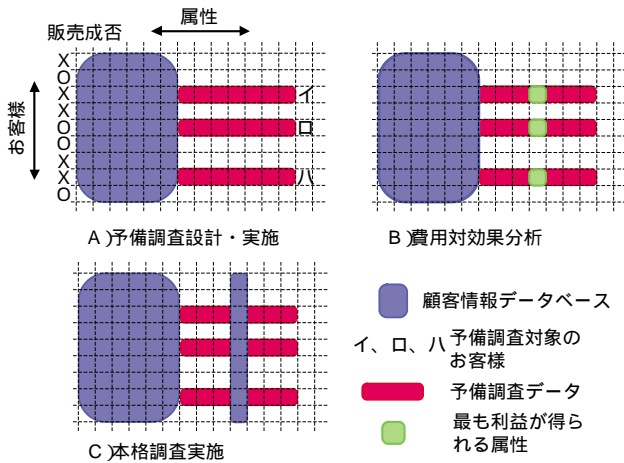


図4-2-4 予備調査/本格調査によるデータベース拡充

## (2) 予備調査設計支援

少数のお客様に対する予備調査によって、追加すべき属性を判断するためには、適切な予備調査を実施する必要があります。不適切な予備調査は、本格調査の実体を捉えそこね、知識の精度向上に貢献しない無駄なデータを収集してしまうリスクを持つ。予備調査のリスクを考慮して、分析者が適切な予備調査を設計(誰に対して、どのような情報を集めるのかを決定)するのは非常に困難である。このような困難を避けるために、システムは以下のような予備調査設計の支援を行う。

予備調査対象の特定: システムはデータマイニングによりデータを分析し、分析結果に基づき、少数の予備調査対象を特定する。

予備調査属性の決定支援: システムは図4-2-6に示すように、現在のデータでは値が非常に似ているため区別ができないお客様の組を予備調査対象として分析

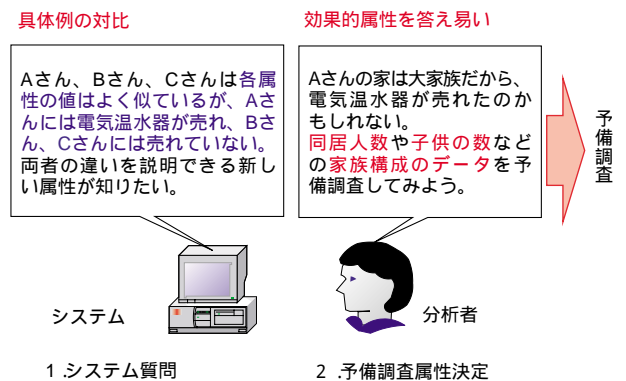


図4-2-6 予備調査属性の聞き出し

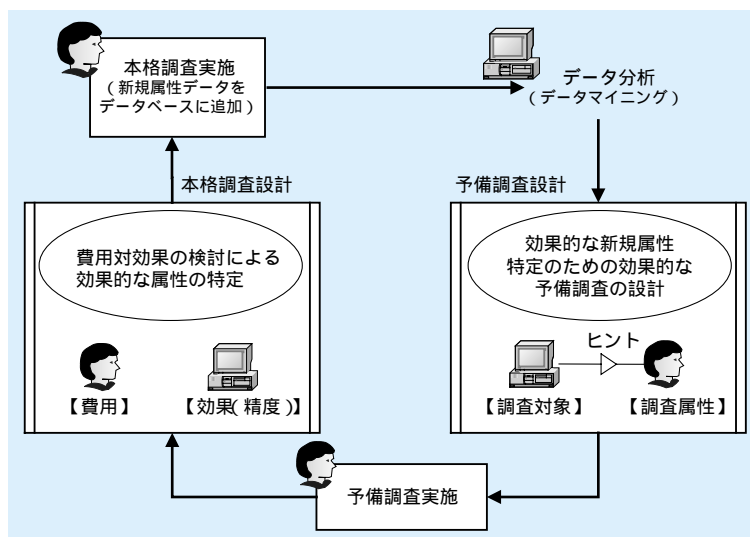


図4-2-5 クリティカル・データ特定法の流れ

者に提示し、お客様を正しく区別できるような属性を聞き出す。このような現在の知識の状況に応じた具体的な対象の提示により、システムが必要としている属性の特徴を分析者が把握することが可能となり、適切な予備調査属性の決定ができる。

さらに、属性追加後のデータベースから抽出される知識の精度の推定値を得るために、本格調査による属性追加後のデータベースからの抽出知識の精度を、非常に少数のお客様への予備調査によって正確に推定するための推定式を導出した。分析者は、各属性のデータ取得コストと抽出知識の精度の推定値を考慮して費用対効果の分析を行い、データ収集コストに見合う精度向上が得られる追加属性を特定する。

### (3) ベンチマークテスト

図4-2-7は、各属性を追加した場合の抽出知識の実際の精度とその推定値の関係を示す実験結果である。適用問題は、5,822人の顧客情報データベースに基づき新たな顧客の販売成否を予測する問題であり、データマイニングの手法の効果を検証するためのデータとして公開されている。図の各点は1つの属性を示し、対応する横軸の値は全5,822人に対する本格調査による属性データ追加後のデータベースからの抽出知識の精度（予測の正解率）を、縦軸の値はシステムが特定した8人に対する予備調査に基づく推定値を示している。図から8人という非常に小規模な予備調査結果に基づいた推定値が実際の精度をうまく推定できていることがわかる。予備調査は本格調査によるデータベース拡充の実体をうまく捉えて

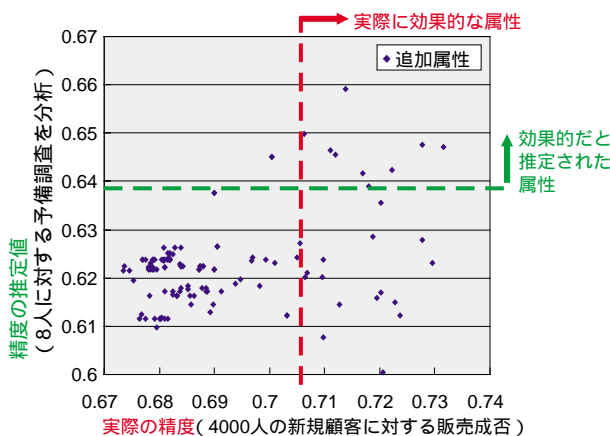


図4-2-7 予備調査に基づく精度の推定値の正確さ

おり、分析者はデータ取得コストとシステムが推定した精度との費用対効果を検討することで、効果的にデータベースを拡充できることがわかる。

### 4-2-3 検索結果の効率的絞込み

#### 背景・目的

近年の情報技術の進歩に伴い、組織および個人で扱える電子化文書の量が急激に増加している。このような状況の中、検索等に習熟していない一般の人が、膨大な電子化文書中から、必要な情報を検索する機会も増加し、必要な情報を効率的に検索できる技術への要望が高まっている。

従来、欲しい文書を検索する場合、求める文書の内容を示す「キーワード」を入力すると、そのキーワードを含む文書リストが得られる「文書検索」が使用される。しかし、利用者の欲しい検索結果（文書リスト）がうまく得られるようなキーワードを指定することは難しい。

当所では、利用者がシステムから提示された文書リストの中から欲しい文書をいくつか指定するだけで、欲しい文書を上手く得られるようなキーワードの組み合わせをシステムが自動的に見つけて、効率的に利用者の欲しい文書リストを得られる非常に効率的な文書絞込み手法を開発した。

#### (1) 基本的考え方

検索において、適切なキーワードを選ぶことは難しい。例えば、ここで述べる「情報フィードバックを用いた文書絞込み」に関して調べたい時、「情報」といった一般的用語をキーワードとすれば、非常に多数の文書が得られ、その中に当然関連文書も含まれるであろうが、無関係な文書も多数得られ、欲しい文書がなかなか見つけれない。一方、「情報、検索、インタラクティブ」といったキーワードで検索をすると、「インタラクティブ」という用語を用いずに「フィードバック」という用語を用いている文書は見つからない。検索に習熟した人は、検索結果を吟味しながら、無関係な文書が少なく、かつ、欲しい文書が極力多数得られるように、検索キーワードを書き換えながら、再検索を繰り返す。すなわち、検索された各文書が欲しい文書であるか否かという判断と、キーワードの修正・再検索という2つの作業を同時に行

っている。しかし、検索に習熟していない利用者にとっては、これは非常に難しい。そこで、利用者は、検索された文書が自分の欲しい文書（適合文書）であるか、否（非適合文書）かの評価に専念し、システムがその評価にもとづいて自動的に検索キーワードの調整を行い、検索結果の絞り込みをするという分業が基本的考え方である。

この考え方は、「適合フィードバック」と呼ばれるが、既存の適合フィードバック手法では、検索精度を効率的に改善できなかった。そこで、サポートベクターマシンとよぶ機械学習手法に基づく能動学習を用いることで、非常に効率的な適合フィードバック手法を開発した。

## (2) 文書検索・絞り込み手順

本手法を適用した文書検索のイメージを図4-2-8に示す。

最初に、利用者は検索キーワードをシステムに入力する。キーワード入力は初回のみである。システムは、そのキーワードに基づいて文書検索を行い、検索結果

の中から、関連性の高い文書n個（予め決めた文書数）を利用者に示す。

利用者は、これらのn個の文書の中から、自分の欲しい文書（「適合文書」）と、明らかに欲しくない文書（「非適合文書」）をいくつか選択評価する。

の利用者評価に基づいて、システムは検索に用いるキーワードとその重みを自動的に調整して再検索を行う。

再検索結果の中から利用者に「適合・非適合」の確認・判断をしてほしいn個の文書を選び出し、利用者に提示する。

利用者が満足できる検索結果が得られるまで数回～を繰り返す。

図4-2-8中、灰色枠内の文書は非適合文書であるが、ユーザが文書の評価すごとに、灰色部分が少なくなり、検索結果の上位は、ユーザの欲する文書が占めるようになる。

ユーザ評価に基づいて利用者の欲しい文書のみが極力得られるキーワードとその重みを決定する（ ）におい

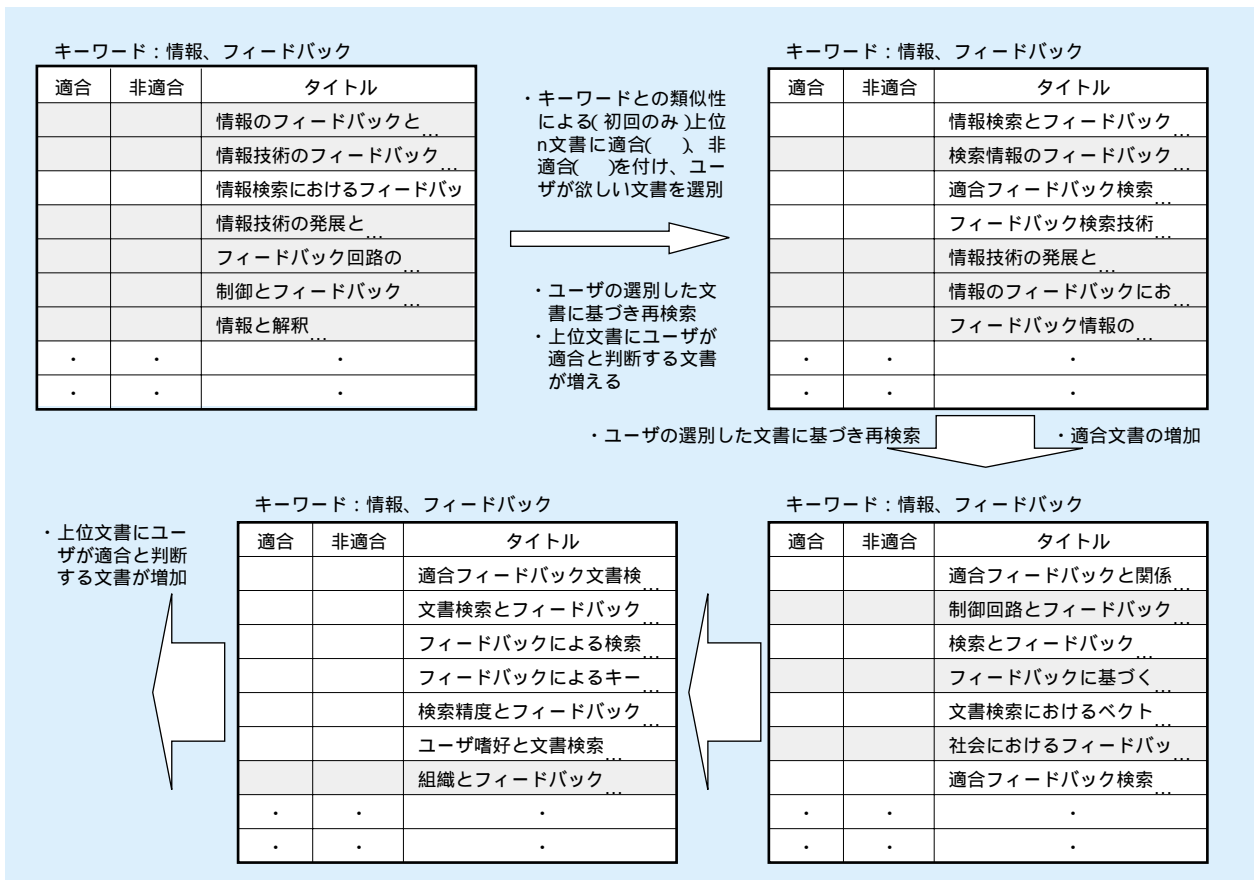


図4-2-8 検索結果の効率的絞り込み

て、サポートベクトルマシンによる機械学習を用いる。さらに、 の利用者に提示する文書の選定では、サポートベクトルマシンの学習結果を用いた各文書の適合の評価推定を用いている。

### (3) 実験結果

文書検索における性能評価で一般的に利用されているベンチマークデータ（The Los Angeles Times 約13万文書、1文書中の平均単語数526語）を用いた実験によって、実用性を検討した。表4-2-1にフィードバック回数と、システムがユーザに提示するために適合と判断した

表4-2-1 フィードバック回数と平均適合率

フィードバック回数	平均適合率	
	従来手法	提案手法
1	23%	26%
2	25%	35%
3	24%	61%
4	23%	64%

上位20文書中（n = 20）の各フィードバック回数における平均適合率を示す。適合率とは、検索された文書の中での適合する文書の割合を表す。

平均適合率が高いことは、ユーザが求める文書が多く提示できることを表し、検索精度が高いことを意味する。表4-2-1より、従来手法では1、2、3、4回とフィードバック回数を増やしても、平均適合率が増えていない。一方、提案手法ではフィードバックの回数を増やしていくに伴って、20文書中にユーザが欲しいと考える文書が増えていることがわかる。これは、たかだか4回のユーザの評価により、ユーザに提示できる20文書中の約13文書を、ユーザの欲する文書で占めることができることを意味している。フィードバック回数がより少なく、ユーザの欲する文書だけをできるだけ多く提示できることが文書検索では重要となる。提案手法では、フィードバック回数4回で、ユーザに提示する文書中に占めるユーザの欲する文書の割合を従来手法の約3倍にでき、文書検索上、実用の域に達していると考えられる。

## コラム5：大量文書自動整理システム「トピックうおっちゃん」

### 文書活用のニーズ

近年、企業は問い合わせ情報や、最新技術動向などの情報を大量に集め、「顧客満足度」や「業務効率」などの向上にすばやく、効果的に役立てようとしている。そのためには収集情報から“最近急増している問い合わせの種類”や“いままでなかった新しい技術情報”など、話題の変化を見つけることが重要である。しかし、このような情報を得るためには、現在までに蓄積された大量の文書情報全体を、情報の内容により分類・整理し、どのような情報がどのくらい増えているのかを調べる必要がある。

### 文書の整理体系の自動発見

当所では、大量の電子文書を自動的に内容の似た文書の集まり（話題）に分類整理する手法により、人手では整理できないほど多くの文書を、同一基準によって短時間で分類し、文書数が増加している話題など、重要と思われる話題を利用者に提示する大量文書自動整理システム「トピックうおっちゃん」を開発した。

「トピックうおっちゃん」による文書整理の最大の特長は、集められている文書全体を分類整理するのに適切な整理体系を自動的に発見し、整理することにある。すなわち、新たな文書が追加されると、既にある文書も含んだ全文書をうまく再分類整理できるような整理体系を新たに作り出す。従来になかった新しい問い合わせが集まるようになれば、それを整理できるような新しい分類（話題）

を作成し、特定の製品についての問い合わせが多くなれば、その製品についてより細かく分類整理するといったことを自動的に行う。このため、時間と共に、収集される文書の内容が変化しても、時々状況に応じた最適な整理を行うと同時に、整理結果である各話題に対して、新しさ、規模、まとまりの良さなど複数の指標を与える。

「トピックうおっちゃん」は、これらの指標を基に、評価の高いものから順に提示する【話題評価・推薦機能】、入力されたキーワードまたは文章に関連が深い話題を検索する【話題検索機能】、話題と関連の深い他の話題や文書を提示する【話題・文書間連携機能】などを有しており、蓄積された文書情報の“変化の迅速な発見”と“有効活用”を支援するシステムとなっている。

### さまざまな文書データベースへ適用可能

「トピックうおっちゃん」は、専門的な単語辞書やその整備を必要とせず、さまざまな分野の文書データベースを1つのシステムで個別に管理することができる。

現在までに、本システムを用いて、当所研究報告書の研究内容別の詳細な整理や、新聞記事、問い合わせ情報、技術情報などさまざまな分野の数万件規模の文書データを整理してきた。そして整理結果や、新しい話題の発見能力や関連情報の質などについて実験評価を行ないその有用性を確認している。

