概念検索に基づく技術内容からのエキスパートの検索

稲子希望[†1,☆] 笠原要[†1] 湯川高志[†2]
加藤恒昭[†3] 北寿郎[†4]

拡張されたベクトル空間法を用い，技術文書とその著者である技術者を同一の多次元空間に配置することで，技術内容の記述に対して，それに精通した技術者であるエキスパートを検索する方式を提案する．キーワードどうしの関連性を考慮した検索が可能であることと，エキスパート検索以外にも多様な検索処理を一貫した枠組みで行えることが特徴である．実験により，従来のベクトル空間法を単純に適用した場合と比べ，提案方式が優位であることを明らかにした．

Expert Recommendation Based on Technical Descriptions Using Concept-based Retrieval

Nozomu Inago,[†1,☆] Kaname Kasahara,[†1] Takashi Yukawa,[†2]
Tsuneaki Kato[†3] and Toshiro Kita[†4]

A method for expert recommendation is proposed. This method processes the description of a technical topic as input and then finds engineers who have a high level of expertise in that area. The technique used is an extended vector space model that can locate both technical topics and engineers in the same multi-dimensional space, and then caluculate their similarity. This method is novel both in the consideration of relationship between keywords and in the implementation of several kinds of retrieval. The advantage of this method over the traditional vector space model was shown in the experiment.

1. はじめに

ナレッジマネジメントは組織が持つ情報を知識に変換し，それと人々とを結び付ける過程であり[1)]，近年，その重要性が認識されると同時に，実現のために様々な情報処理技術の適用が検討されている[2)]．ナレッジマネジメントの1つの側面として，人と人とを結び付けることがあげられる．これは，巨大化し合併等で流動化の激しい組織において，古き良き「人脈」が持っていた役割をシステムに代替させようという試みである．たとえば，以下のような状況での利用が想定される．

- 新聞記事の中にライバル会社の技術発表をみつけて，その技術に関する詳細や自社の開発状況を把握したいマネージャに対して，自社内のだれに問い合わせればよいかをアドバイスする．
- 前例のないトラブル報告を受け取ったヘルプデスク担当に対して，その種の問題に詳しい技術者を紹介する．

本稿では，話題となっている技術内容や問題に対して，その解説や解決にふさわしい専門的能力を有する人物（エキスパート）を紹介する方式を提案する．本提案でのアプローチは，ある人物の専門的能力をその人物が執筆した文書，たとえば技術文書や事例報告書の内容を用いて推定して蓄積しておき，それと問い合わされた技術内容，つまり技術発表記事やトラブル報告で述べられている内容との類似度を計算することで，その技術内容にふさわしい人物を紹介するというものである．

まず，2章で本提案の概要を述べる．次に，3章でエキスパート検索タスクにおける評価実験について説

†1 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corporation
†2 長岡技術科学大学工学部電気系
Department of Electrical Engineering, Nagaoka University of Technology
†3 東京大学大学院総合文化研究科
Graduate School of Arts and Science, The University of Tokyo
†4 NTT コミュニケーションズ株式会社
NTT Communications Corporation
☆ 現在，西日本電信電話株式会社
Presently with NTT West Corporation

明し，方式の有効性を検証する．4 章で本提案の特徴について考察する．

2. 提案方式の概要

文書や質問を多次元空間のベクトルとして表現し，それらのベクトルのなす角の余弦係数を類似度として情報検索を行う方法は，ベクトル空間法としてよく知られている[3]．基本的な手法では，この多次元空間の軸にキーワードを対応づけ，文書におけるあるキーワードの重みをその軸の座標としている．キーワードと軸とを直接対応づけて得られる空間を直交空間として扱うため，同義語の存在やキーワードの出現共起等の現象，すなわち，キーワードどうしの関連性がまったく配慮されないことが問題として指摘されている．

この問題に対して 2 つの方法が提案されている．Deerwester らによる Latent Semantic Indexing（LSI）は，キーワードと文書の集まりに対して，どのキーワードがどの文書に何回出現したかという出現頻度行列を作成し，特異値分解（Singular Value Decomposition，SVD）を用いてそれを次元圧縮することにより，キーワードと文書が配置された多次元空間を得る[4]．Schütze らによる共起によるシソーラスとその情報検索への応用では，与えられた文書集合内での単語の近傍共起に基づいてキーワードを多次元空間に配置し，文書をその中に現れるキーワードの出現頻度（tf）を考慮した重心へと配置する[5]．笠原らは計算機可読の国語辞書を用いて単語を多次元空間に配置する方法を提案しており[6]，そのようにして得られたキーワードベクトルに基づいて Schütze らと同様に文書をそこに現れる単語の重心として表現する方法も検討されている[7]．これらの手法は，質問を構成するキーワードを文書が含んでいることを必ずしもそれが検索されるための条件とせず，キーワードどうしの意味的な関連性を考慮した検索であるため，概念検索と呼ばれることがある[8]．

概念検索で重要なことは，キーワードと文書というある意味ではまったく異質なものを，類似度が定義できる同じ 1 つの多次元空間に配置したことである．このような配置により，キーワードどうしの類似性，キーワードと文書との類似性，文書どうしの類似性が定義でき，それによって，同義語や類義語等のキーワードどうしの関連性の問題，キーワードを条件とする一般の文書検索の問題，そして適合性フィードバックや類似文書検索の問題を，多次元空間における類似度計算という統一的な枠組みで扱うことが可能となっている．

本稿では，これを一歩進めて，キーワードや文書だけでなく，それを執筆した技術者も同一の多次元空間に配置することを試みる．文書のベクトル $\overrightarrow{D_j}$ や技術者のベクトル $\overrightarrow{A_k}$ をキーワードベクトル $\overrightarrow{w_i}$ を用いて以下のように表現する．

$$\overrightarrow{D_j} = \sum_i c_{ij} \overrightarrow{w_i} \tag{1}$$

$$\overrightarrow{A_k} = \sum_{j \in Pub(k)} \frac{\overrightarrow{D_j}}{|\overrightarrow{D_j}|} \tag{2}$$

ここで，c_{ij} はキーワード i の文書 j での重みで，後述するように出現頻度である tf 値，あるいは tf · idf 値を用いる．$Pub(k)$ は技術者 k が著者となっている文書の集合である．質問についても，式 (1) を用いてそのベクトル表現を得る．

このような表現により，文章やキーワードのならびで表現された質問，文書そのもの，その執筆者である技術者を同一空間内に配置し，ベクトルの余弦係数を用いて，それらについて相互に類似度の計算を行うことができる．話題となっている技術内容や問題の表現を構成するキーワードの重みつき和として得られるベクトル表現と，執筆した文書を構成するキーワードの重みつき和として得られる技術者のベクトル表現とを比較し，より類似しているベクトルで表現されている技術者ほど，その解説や解決にふさわしい専門的能力を有するエキスパートと考えるのである．

単純な実現であるが，文書やその著者である技術者をキーワードと同一の多次元空間に配置し，種類が異なるもの（キーワード，文書，技術者）どうしの類似度計算を可能としたことは重要で，以下の処理が可能になる．

- 多様な検索
  キーワードのならび，あるいは技術内容を記述した文章（文章はキーワードの集まり（bag）と見なす）を質問として，その技術に関するエキスパートである技術者を検索できる．加えて，逆にある技術者の関心の近い文書を検索する，ある技術者と専門的興味が近い技術者を検索する，技術内容を条件にそれと内容が近い文書を検索する等々が可能である．これら多様な検索処理を一貫した枠組みで行うことができる．
- 自動分類
  検索されたエキスパートの集団を，クラスタリングの手法を用いて自動的に分類することができる．さらに，得られた分類を特徴づけるキーワード（や文書）を自動的に付与することができる．つまり，scatter-gather[9]で文書について行われ

ていた検索をそのままエキスパートの検索に置き換えた処理が可能である．検索されたエキスパートの集団に対して，利用者がいくつかのキーワードを与え，それらとの関連を基準として集団を分類する等の高度な処理も可能である．

このように，概念検索によるエキスパート検索には様々な長所があげられる．しかし，エキスパート検索自体はタスクとしてこれまでに十分検証されていないため，どのような種類のキーワードベクトルを利用するか，文書を表現するときの単語のキーワードベクトルの重み（式 (1) の c_{ij}）としては tf と $\mathrm{tf}\cdot\mathrm{idf}$ のどちらが有効であるか等については，実験的に明らかにすることが必要である．また，古典的なベクトル空間法で表現された文書ベクトルから技術者のベクトルを式 (2) に準じて作成すれば，キーワードどうしの関連性が考慮されていない等の欠点があるが，形式的にはエキスパート検索が同様に可能となる．そこで次章では，提案方式の有効性を明らかにするために行った評価実験について説明する．

3. 評価実験

本稿では，提案方式の最も基本的な機能である文書からのエキスパート検索について，特許文書を用いて評価する．提案方式のその他の機能の実例やそれを用いたインタラクションについては，文献 10) に詳しいので，そちらを参照いただきたい．

3.1 テストコレクション

情報検索において適合率，再現率等の評価値を計算するためには，質問とそれに対する正解がセットになったテストコレクションが必要である．文書検索に対しては NTCIR-1☆のようなテストコレクションが存在するが，エキスパート検索に対するテストコレクションは存在しない．そのため，対象とする専門分野を設定し，いくつかの技術内容（質問）に対するエキスパートとしてふさわしい技術者をその分野の専門家に列挙してもらうことにより，テストコレクションを作成することが方式評価のために必要である．コレクションの信頼性を高めるためには，複数人の専門家に列挙してもらう必要があるが，そのためには結局，専門家の選定を先に行うことになるので，作成は本質的に難しい．また，実際的な側面でも，テストコレクションの作成は多くの時間と人手を要する．そこで，エキスパート検索を近似的に評価するためのテストコレクションを以下の方法で自動作成した．

評価用の文書集合として，平成 11 年公開特許広報を用いた．これには，平成 11 年に公開された特許文書 38 万件が収録されており，分野ごとに分類されている．38 万件という膨大な量の文書すべてを用いた評価実験は困難なため，その一部の文書集合を用いた．特許の分類構造は IPC（国際特許分類）と呼ばれる木構造になっており，根の側から順に，セクション，クラス，サブクラスといった階層（カテゴリ）がある．本稿では，サブクラス G06F（「電気的デジタルデータ処理」）に含まれる 2 万 3 千文書とその著者（技術者）2 万 9 千人を用いてテストコレクションを作成した．共著もあるため，技術者 1 人あたりの文書数は約 2.1 文書であった．サブクラスは，さらにメイングループ，サブグループへ階層的に分類されている．同じサブグループ内の特許どうしは同じ分野の特許と判定されたものであり，ある程度技術内容も近いと考えられる．そこで，同じサブグループ内の特許を執筆した技術者をそのサブグループの特許内容に精通したエキスパートと見なした．

そして，質問として選択された文書（以下，質問文書）に対し，その質問文書と同じサブグループに含まれる文書の著者を適合するエキスパート，それ以外を不正解とした．質問文書 100 件を対象の G06F サブクラスからランダムに選択し，それ以外の文書（以下，検索対象文書）を用いて式 (2) の技術者ベクトルを作成した．

エキスパート検索の質問としては，質問文書中の「要約」と「発明の名称」（以下，タイトル）を個別に用いた．どちらも特許の概要を表現したものだが，タイトルは要約に比べて短く，含まれる単語が少ないため，異なる種類の質問における提案方式の有効性を検証できる．今回作成したテストコレクションの質問文書では，要約の平均単語数が約 167，タイトルが約 7 であった．質問文書の要約またはタイトルを質問として入力し，検索結果を適合率，再現率で評価した．

3.2 技術者ベクトルの作成

概念検索として，キーワードのベクトルをまず作成し，それを用いて文書ベクトルを作成する手法を用いる．本稿では，これを拡張されたベクトル空間モデル（以下，EVSM☆☆）と呼ぶ．

EVSM におけるキーワードベクトル $\overrightarrow{w_i}$ の作成方法として，コーパスを用いる方法[5]と国語辞書を用いる方法[6]を比較した．

- **コーパスを用いたキーワードベクトルの作成方法**

☆ NACSIS Test Collection for IR System 1

☆☆ Extended Vector Space Model

コーパスに含まれるキーワードに対し，その周辺に現れる単語（共起語）を属性として抽出して，共起語を軸とする多次元空間にキーワードを配置する方法．具体的には，キーワードと同じ文中に現れる単語を共起語として用いる．

- **国語辞書を用いたキーワードベクトルの作成方法** 国語辞書の見出語（キーワード）に対し，その説明文中に現れる単語（説明語）を属性として抽出して，説明語を軸とする多次元空間にキーワードを配置する方法．

コーパスを用いる方法は単語の使われ方に基づいて，一方，国語辞書を用いる方法は単語の語義に基づいて単語間の関連性を定義するものだといえる．以後，コーパスによるキーワードベクトルを利用したEVSMを**コーパス型EVSM**，国語辞書によるものを**辞書型EVSM**と呼ぶ．一般的な文書検索においては，コーパス型EVSMと辞書型EVSMによる手法は正解の傾向が異なるものとなっていると報告されているので[7]，エキスパート検索における有効なキーワードベクトルを明らかにするために比較した．

エキスパート検索における概念検索そのものの有効性を確認するために，古典的なベクトル空間モデル（以下，VSM☆）についても，評価を行った．技術者が著作した文書のベクトル和でその技術者のベクトルを表現する点では，EVSMにおける技術者のベクトル作成方法と同じである．しかしVSMにおける文書のベクトルの要素は，文書中のキーワードの重みそのもの（式(1)のc_{ij}）であり，EVSMによる文書ベクトルと異なる．

さらに，2種類のEVSMとVSMそれぞれで文書を表現するためのキーワードの重みとして，出現頻度であるtf値を用いた場合と，tf·idf値を用いた場合を比較した．

なお，以下の評価実験では，形態素解析として日英翻訳システムALT-J/E[11]の日本語形態素解析部を利用した．

具体的な手法としては，まず，VSMに関しては，式(3)（tf値を用いる場合），式(4)（tf·idf値を用いる場合）を使って文書ベクトルを作成し，それを基に式(2)の技術者ベクトルを作成した．ここで，tf_{ij}はキーワードiの文書jでの出現頻度である．また，idf_iは文書集合に対するキーワードiのidf値であり，文書数をN，キーワードiを含む文書数をdf_iとすると，$idf_i = \log N/df_i$である．

☆ Vector Space Model

表1 評価実験のパラメータ
Table 1 parameters.

パラメータ	パラメータのとる値
モデル	[コーパス型EVSM], [辞書型EVSM], [VSM]
文書ベクトル作成時のキーワードに対する重みづけ	[tf], [tf·idf]
質問として入力する文章	[要約], [タイトル]

$$\overrightarrow{D_j} = (tf_{1j}, \ldots, tf_{nj}) \tag{3}$$

$$\overrightarrow{D_j} = (tf_{1j}idf_1, \ldots, tf_{nj}idf_n) \tag{4}$$

コーパス型EVSM，辞書型EVSMについては，式(1)，式(2)を用いて，文書ベクトル，技術者ベクトルを作成したが，キーワードベクトル$\overrightarrow{w_i}$の作成はそれぞれ以下のように行った．

コーパス型EVSMにおけるキーワードベクトルは文献5)を参考に，以下のようにして作成した．検索対象文書の要約とタイトルの集合をコーパスとし，出現頻度が高い自立語2万語をキーワード，その中でも頻出の1,000語を多次元空間の軸に対応させる単語（以下，属性語）として選択した．この2万語のキーワードに対し，それと同じ文中に出現する属性語の共起頻度を要素する単語–単語の共起頻度行列を作成した．キーワードと属性語を出現頻度が高い語に絞ったのは，頻度が低い共起からは統計的に信頼性の高いキーワードベクトルが作成できないと考えたためである．最後に，共起頻度行列に対してSVDにより次元を圧縮した．コーパス型EVSMにおけるキーワードベクトルの適切な次元数については，文献12)で検討されており，その中では最適と判断された100次元を本稿でも同様に圧縮する次元数とした．

辞書型EVSMにおけるキーワードベクトルは文献6)にならい，以下のようにして作成した．国語辞書としては学研国語大辞典[13]を利用した．まず，9万語の見出語（キーワード）に対する説明文中の単語を属性語として抽出し，その頻度を要素とする単語–単語の行列を得た．次に，見出語と属性語，見出語どうしの関連性を考慮して行列を精錬した．実際には，得られた行列に，その転置行列や2乗した行列を線形結合し，idfで値を補正した．最後に，属性語を日本語語彙大系[11]のシソーラスの3千カテゴリで一般化することにより，3千次元のキーワードベクトルを得た．詳しくは文献6)を参照されたい．

評価実験のパラメータを**表1**にまとめる．6種類の方式（モデル3種類 × 重みづけ2種類）に対して，

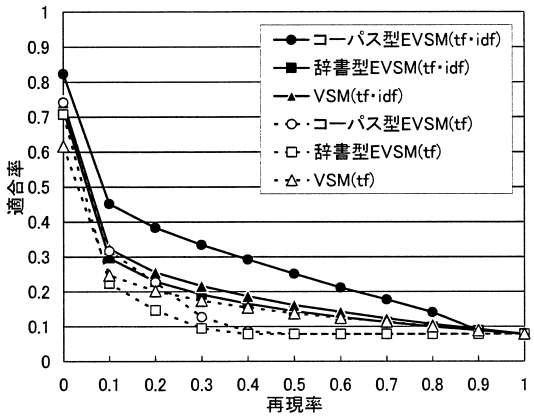


図 1 実験結果（質問：要約）
Fig. 1 Result (query: abstract).

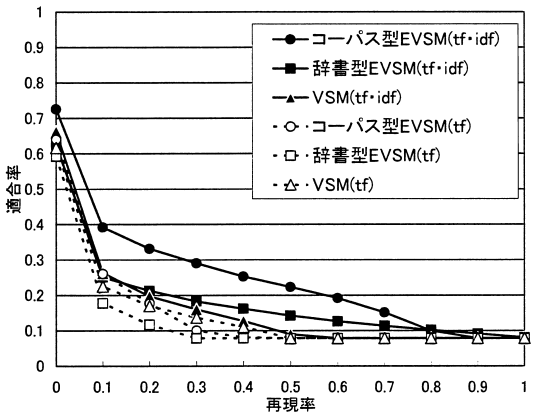


図 2 実験結果（質問：タイトル）
Fig. 2 Result (query: title).

2 種類の入力を与えてエキスパート検索を行い，評価した．

3.3 実験結果

特許の要約を質問としたときの実験結果を**図 1**，タイトルを質問としたときの実験結果を**図 2** に示す．

図 1，図 2 のどちらにおいてもコーパス型 EVSM と tf·idf による方式が最も高い適合率となっている．全体的にタイトルを入力したとき（図 2）の方が適合率が低く，要約を入力したとき（図 1）よりも検索が困難であることが分かるが，各方式間の相対的な関係はほぼ同じである．

最も良い結果を出している要約を入力とした場合のコーパス型 EVSM 方式で適合率と再現率が一致するのは，およそ 0.3 のあたりである．正確さを欠く表現であるが，これは，10 人のエキスパートが存在するような分野を対象に，10 人の技術者を検索したときに，そのうちの 3 人が真のエキスパートであることを示している．また，再現率 0.1 のときの適合率は 0.5 付近であるので，同じ分野で 2 人の技術者を検索した場合，そのどちらかは真のエキスパートである．エキスパート検索にどの程度の適合率と再現率が必要とされるかについて，客観的な基準は存在しないが，このような数値であれば，少なくとも，「はじめに」で述べた状況で，十分に有効なツールとなると考えられる．

モデルを比較すると，重みづけが tf のときはモデル間の差はあまりないが，tf·idf のときはコーパス型 EVSM が他の 2 つのモデルよりも飛躍的に高い適合率となっている．また，辞書型 EVSM においても，VSM と同程度以上の適合率となっており，EVSM はエキスパート検索において有効であるといえる．もちろん，キーワードの多次元空間への配置を闇雲に行っていては EVSM はうまく働かない．本実験の結果は，キーワードの関連性を反映した配置ができていることを示している．辞書型 EVSM よりもコーパス型 EVSM の方が高い適合率となっているのは，コーパス型 EVSM では検索対象文書そのものを情報源としてキーワードを多次元空間に配置するので，検索対象文書内でのキーワードどうしの関連性をより適切に反映しているためと考えられる．

また，辞書型 EVSM では，国語辞書に含まれない文書中の単語はキーワードベクトルがないために文書ベクトル作成に利用できない．たとえば専門用語は技術者を特徴づける単語として重要と考えられるが，そのような単語は国語辞書に含まれないことが多い．形態素解析辞書においても未知語となっている場合があるが，コーパス型 EVSM では未知語もキーワードとして利用しているので，これも辞書型 EVSM より適合率が高い理由の 1 つと考えられる．ただし，逆に国語辞書にはコーパスに含まれない非常に多くの単語が含まれており，そのような単語は文書ベクトルを作成するときには利用できないが，ユーザが質問を入力するときにはその語彙を選ばないという意味で有用である．ユーザが専門用語等を知らない場合に特に有効であると考えられ，これは VSM やコーパス型 EVSM にはない利点である．また，コーパス型 EVSM におけるキーワードベクトルの確からしさはコーパスの規模に依存すると予想される．定量的な評価は行っていないが，数百件程度の文書について同モデルを適用する場合には，適切なエキスパート検索が行えない恐れがある．それに対して辞書型 EVSM のキーワードベクトルは，検索対象の文書の質や量には依存しないので，エキスパート検索システムに本方式を適用する初期段階で文書が十分収集できないような場合に有効と予想される．

重みづけの比較をすると，EVSM，VSM のいずれも tf·idf の方が tf よりも高い適合率となっており，一般的な文書検索と同様に idf による tf の補正はエキスパート検索においても有効であると予想される．また，idf の有効性は，コーパス型 EVSM，辞書型 EVSM において顕著である．EVSM では，キーワードベクトルを線形結合する際の結合定数の補正に idf を用いており，拡張したベクトル空間法において idf が本質的に有効な補正であることを示唆している．

文書検索で VSM と EVSM を比較した文献 7) においては，EVSM を用いると VSM では不可能な検索が可能であるが，全体の精度としては VSM の方が高くなったと報告されている．ただし，文献 7) における EVSM は idf が考慮されていない．本稿のエキスパート検索において示されているように，文書検索においても idf を考慮することによって EVSM の精度が高くなる可能性が高い．

今回の実験では，テストコレクションを自動作成するために，文書があらかじめ分類されている特許文書を利用したが，方式を実装するために分類情報や学習データは必要としない．そのため，他の種類の文書，たとえば分類情報が付与されていない論文データベースを用いたエキスパート検索も可能である．

3.4 実 行 例

評価実験で最も高い適合率となった tf·idf とコーパス型 EVSM を用いたエキスパート検索の実行例を示す．ある特許文書中の一文である「テキストに含まれる振り仮名の量を利用者のレベルに応じて適切に調整して生成し，読みやすさを向上させた振り仮名制御装置を提供する.」という文章を質問として入力した．その結果，上位 3 位までの技術者はこの特許の著者である IT 氏，TH 氏，MM 氏であった．4 位の IN 氏と 5 位の EE 氏はいずれも漢字の振り仮名の付与に関する特許を執筆している．ここで注目しておきたいのは，EE 氏が執筆した特許の中では,「振り仮名」ではなく「ルビ」という単語を使っている点である.「振り仮名」という単語は使っていないが，それに意味が近い「ルビ」という単語を使っているために EE 氏を上位に検索しており，EVSM の効果が確認できる．6 位の IM 氏と 7 位の SK 氏も振り仮名の付与に関する特許を多数執筆しているが,「振り仮名」ではなく「読み」という言葉を使っており，同様の効果がうかがえる．

4. 考 察

4.1 既存技術との比較

本方式の特徴の 1 つは，文書やその著者である技術者を同一の多次元空間に配置し，種類が異なるもの（質問，文書，技術者）どうしでも類似度計算を可能としたことである．これによって，技術内容からのエキスパート検索だけでなく，たとえば，ある技術者の関心の近い文書を検索する，ある技術者と専門的興味が近い技術者を検索する，技術内容を条件にそれに内容が近い文書を検索する等々，多様な検索処理を一貫した枠組みで行うことができる．

EVSM を用いる方式では，文書や技術者だけでなくキーワードも同じ空間に配置しており，同様に類似度計算の対象とできる．これによって，ユーザが選択したいくつかのキーワードをクラスとし，文書や技術者を最も類似度が高いクラス（キーワード）に割り当てて自動分類を行ったり，文書や技術者と類似度が高いキーワードを検索して，それらを特徴づけるキーワードを自動的に付与したりするといった処理が可能である．たとえば，3.4 節の実行例で 1 位に検索された IT 氏に対しては,「難読」,「振り仮名」,「テキスト」,「熟語」等が IT 氏を特徴づけるキーワードとして付与された．もちろん EVSM の最大の特徴は，キーワードベクトルをもとに文書や技術者を表現することにより，キーワードどうしの関連性を考慮した検索が可能なことである．また，EVSM は VSM が持っている利点をそのまま継承しており，クラスタリングや適合性フィードバック等，従来の VSM で可能な多くの処理は EVSM においても適用可能である．

ブーリアン検索を用いて人物検索を行うシステム[14]も存在するが，ブーリアン検索は検索処理が高速であるという利点はあるものの，キーワードの有無に関する条件の解釈が厳しいので，必要となる技術内容について詳細に指定することは容易ではない．VSM や EVSM では，少数のキーワードだけでなく，もっと長い文章，たとえば，新聞記事や事例報告全体を条件とすることができる．このため単なるトピック分野ではなく，より詳細な技術内容についての質問が可能である．このことは文書検索でもあてはまるが，指定されるキーワードの数が比較的少ないといわれる文書検索に比べ，本方式が利用される状況を考えると，その利点はより大きな価値を持つと考えられる．

ほかにも，本方式と同じように人物を紹介するシステムとして ContactFinder[15]がある．これは掲示板にあげられた問合せに対して適当な人物を紹介するシステムだが，その背景となる技術はヒューリスティクスに基づくキーワード抽出による文書のトピック分野推定である．したがって，拡張されたベクトル空間法を用いる本方式とは基本技術が異なり，前述したよう

な本方式の利点は当然持っていない．

4.2 本方式の拡張

本方式では，技術者をその技術者が執筆した文書のベクトルの重心として表現した．これは，文書に付与されている著者情報を利用したものだが，文書には著者である技術者名だけでなく，その技術者が所属する組織名も付与されている場合が多い．この組織についても技術者と同様に多次元空間に配置することを考える．やり方は文書のベクトルから技術者ベクトルを作成する式 (2) の応用で，組織をそこに所属する技術者のベクトルの重心として定義する．部署名と企業名のように，組織の情報が階層的に記述されている文書の集合もあるだろう．この場合は技術者ベクトルから部署のベクトルを作成し，部署のベクトルから企業のベクトルを作成する．組織の情報がより多くの多層構造になっている場合もやり方は同様である．逆に，著者が書かれておらず，発行元の組織名のみが書かれている文書の集合もあるだろう．その場合は文書のベクトルから直接，組織のベクトルを作成する．このように，文書に付与されている著者以外の情報（組織等）も多次元空間に配置し，質問や検索対象となるように本方式を拡張することが可能である．

4.3 エキスパートの定義

本方式では余弦係数を類似度としている．技術者同士の関心や専門性の類似という点では，その技術者が執筆した文書の数は問題にならないだろう．その点では余弦係数は適切な尺度といえる．しかし，ただ 1 つの文書を執筆した人物よりも関連した主題で数十本の文書を執筆している人物の方がエキスパートとして信頼にたると考えることもできる．もし執筆文書数を考慮したいのであれば，たとえば類似度を内積とすることによって類似度計算に反映させることは可能である．しかし，専門性が執筆文書数に比例するのかその対数に比例するのかの選択等については，恣意的になることは免れない．

本稿の評価実験では，ある分野の特許を 1 本でも執筆していれば，その分野のエキスパート（正解技術者）であると見なしているが，ある分野のエキスパートとはいったいどういう人物かを厳密に定義することは実は難しい．特許文書を対象とする場合は，このような問題は現れにくいが，論文等を対象とすると，たとえば「21 世紀の自然言語処理」という解説記事を 1 本だけ執筆している人物と，統語解析から意味解析，文脈処理に及ぶまでの専門的研究論文を多数執筆している人物とを比較して，「自然言語処理に詳しいのどちらか」という問いに答えるのは容易ではない．たとえば EVSM では，執筆文書数は考慮していないことと，統語解析，意味解析，文脈処理に関連するキーワードの和が必ずしも「自然言語処理」のベクトルと一致しないことから，前者を上位に検索すると思われる．解説記事の執筆が幅広く深い専門性を反映していることは疑いないので，質問の性格上，これを適切であると考える人と，納得できない人とに分かれるのではないかと思われる．この問題に対しては，分野の階層関係を考慮することで何らかの改善もしくは変化が期待できる．実は，この問題は EVSM におけるキーワードのレベルにまで投影されるもので，上位下位関係にあるキーワードが多次元空間でどのように配置されるべきかは興味深い問題である．

4.4 技術者の専門性の表現

本方式では，1 人の技術者はただ 1 つの専門分野を持つという仮定のもと，1 人の技術者を 1 つのベクトルで表現している．しかし，1 人の技術者が複数の専門分野を持つことも考えられる．特に，一定以上の期間にわたって執筆された文書を集めた場合はその感が強い．たとえば，本方式では，機械翻訳技術と情報検索技術の両方で特許を執筆している技術者のベクトルがその中間に位置してしまい，どちらの分野でもエキスパートとして検索されないことが起こりうる．また，3.4 節のように，本方式を組織，企業へと拡張した場合はさらにその専門性が曖昧になると思われる．

Schütze らは，ある語が現れる文脈を多次元空間ベクトルとして表現し，それをクラスタリングすることで文脈をいくつかに分類し，それぞれの語の現れの文脈がどの分類に属するかで多義語の曖昧性解消を行っている[16]が，同じ枠組みで，複数の専門分野を持つ技術者に対処することが考えられる．つまり，ある技術者が執筆した文書を一定の閾値でクラスタリングして，得られたクラスタの数だけの専門性，専門分野を持つとするのである．組織や企業についても同様に，そこに含まれる文書や技術者をクラスタリングすることができる．あまり細かいクラスタリングをしてしまうと，1 つの文書が 1 つの専門分野に対応することになってしまう．適当な閾値が設定できるかは興味深い問題である．

5. おわりに

技術文書とその著者である技術者を同一の多次元空間に配置することで，技術内容の記述に対して，それに精通した技術者であるエキスパートを検索する方式を提案した．従来のベクトル空間モデル（VSM）と，キーワードの関連性を考慮した検索が可能である拡張

されたベクトル空間モデル（EVSM）による方式を実装した．評価実験を行ったところ，コーパスにおけるキーワードの関連性を利用したEVSMがエキスパート検索に最も有効であることが分かった．

今後は，ある技術内容に対するエキスパートとはどんな人物であるか，技術者の専門性をどのように表現するのが適切であるかを熟慮し，それを方式に反映させる方法論については，評価方法とあわせて検討する予定である．また，概念検索と同様に，質問に直接出現しないが関連するキーワードを含む文書を検索できる方式として，検索結果の上位の文書で質問を拡張する擬似関連フィードバックがあり，これをエキスパート検索に適用した場合の本方式との比較についても検討を加えることを予定している．

参 考 文 献

1) O'Leary, D.: Knowledge-Management Systems: Converting and Connecting, *IEEE*, Vol.13, No.3, pp.30–33 (1998).
2) 人工知能学会：合同研究会 "AIシンポジウム'99"（第10回）ナレッジマネジメントとその支援技術, SIG-J-9901 (1999).
3) Salton, G. and Buckley, C.: Term weighting approaches in automatic text retreival, *Information Processing & Management*, Vol.24, No.5, pp.513–523 (1988).
4) Deerwester, S., Dumais, S., Furnas, G., et al.: Indexing by Latent Semantic Analysis, *J. Am. Soc. Inf. Sci.*, Vol.41, No.6, pp.391–407 (1990).
5) Schütze, H. and Pedersen, J.: A cooccurrence-based thesaurus and two applicaions to information retrieval, *RIAO'94*, pp.307–318 (1994).
6) 笠原 要，松澤和光，石川 勉：国語辞書を利用した日常語の類似性判別，情報処理学会論文誌, Vol.38, No.7, pp.1272–1284 (1997).
7) 熊本 睦，島田茂夫，加藤恒昭：概念ベースの情報検索への適用—概念ベースを用いた検索の特性評価—，情報処理学会研究報告，ICS-115, pp.9–16 (1999).
8) 日経BP：概念検索が可能な検索エンジン，日経オープンシステム，No.67, pp.108–111 (1998).
9) Cutting, D., Karger. D., Pedersen. J. and Tukey. J.: Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, *ACM SIG-IR'92*, pp.318–329 (1992).
10) 加藤恒昭，笠原 要，北 寿郎：概念検索に基づく技術内容からのエキスパートの推定，電子情報通信学会技術研究報告，NLC2000-8, pp.55–62 (2000).
11) 池原 悟，宮崎正弘，白井 諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林 良彦：日本語語彙大系，岩波書店 (1997).
12) Schütze, H.: Dimensions of Meaning, *Proc. Supercomputing 92*, pp.787–796 (1992).
13) 金田一春彦，池田弥三朗：学研 国語大辞典 第二版，学習研究社 (1988).
14) 榎本俊文，牛島浩一，佐藤哲司：NTTグループにおけるナレッジマネジメント KnowWho 検索を活用した研究開発情報に関するナレッジマネジメント，NTT技術ジャーナル，Vol.12, No.5, pp.24–26 (2000).
15) Krulwich, B. and Burkey, C.: The ContactFinder agent: Answering bulletin board questions with referrals, *AAAI-96*, Vol.1, pp.10–15 (1996).
16) Schütze, H. and Pedersen, J.: Information retrieval based on word senses, *4th Annual Sympo. on Document Analysis and Information Retrieval*, pp.161–175 (1995).
17) Yukawa, T., Kasahara, K., Kato, T. and Kita, T.: An Expert Recommendation System using Concept-based Relevance Discernment, *Proc. 13th International Conference on Tools with Artificial Intelligence*, pp.257–264 (2001).

(平成15年 2 月17日受付)
(平成15年11月 4 日採録)

稲子 希望（正会員）

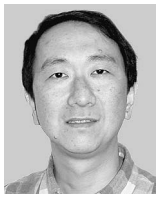
1998年九州大学大学院システム情報科学研究科情報理学専攻修士課程修了．同年に日本電信電話株式会社に入社後，2003年6月まで大規模知識ベースの研究に従事．現在，西日本電信電話株式会社に所属．

笠原 要（正会員）

1991年東京工業大学大学院総合理工学研究科電子化学専攻修士課程修了．同年日本電信電話株式会社に入社．知識処理技術，特に大規模知識ベースの研究に従事．現在，NTTコミュニケーション科学基礎研究所研究主任．1998年11月より1999年11月までスタンフォード大学CSLI滞在研究員．1998年人工知能学会奨励賞受賞．人工知能学会，言語処理学会各会員．

湯川　高志（正会員）

1987 年長岡技術科学大学大学院工学研究科電気電子システム工学専攻修了．同年，日本電信電話株式会社入社．1995 年 1 月より 1996 年 1 月まで，米国南メソジスト大学客員研究員，2002 年より長岡技術科学大学電気系助教授，現在に至る．人工知能向け並列コンピュータ，柔らかい推論システム，テキストに基づく知識処理システムの研究に従事．博士（情報学）．電子情報通信学会，人工知能学会，IEEE 各会員．



加藤　恒昭（正会員）

1981 年東京工業大学工学部電気電子工学科卒業．1983 年東京工業大学大学院総合理工学研究科電子システム専攻修士課程修了．同年，日本電信電話公社（現 NTT）に入社．2000 年より，東京大学大学院総合文化研究科言語情報科学専攻助教授，現在に至る．自然言語理解，対話処理，マルチモーダルコミュニケーションに関する研究に従事．工学博士．電子情報通信学会，言語処理学会，ACL 各会員．

北　　寿郎

1976 年名古屋大学大学院機械工学専攻修士課程修了．同年，日本電信電話公社（現 NTT）に入社．メカトロニクス，ロボティックス，コミュニケーション科学に関する研究に従事．現在，NTT コミュニケーションズ株式会社 J プラットフォーム推進室長．工学博士．電子情報通信学会，日本機械学会各会員．