共起単語間の関連性を考慮した文書重要度付与

高木 徹[†] 木谷 強[††]

大規模な文書データベースを対象とするフルテキスト検索では検索ヒット件数が多くなる傾向があるため，ユーザの検索作業を支援する観点から，検索結果に対して重要度を付与する必要がある．本論文では，重要度算出手法として文書内における検索語の出現共起情報を用いる手法を提案する．単語の共起情報として，近接出現距離，共起検索語間の関連性，および共起検索語の重要度を用いて共起重要度を算出し，単語頻度情報から得られる文書の重要度と組み合わせて重要度を算出する．日本語の情報検索評価用テストコレクションを使用し，単語頻度情報のみによる重要度付与手法と，共起情報を考慮した提案手法の検索精度を比較した．この結果，提案手法の平均適合率が約 0.098 向上 (従来手法と比較した場合の向上率 37%) することを確認した．

Relevance Ranking of Documents Using Query Word Co-occurrences

Toru Takaki[†] and Tsuyoshi Kitani[††]

Full text search from huge databases tend to give a great number of retrieved documents. To help user's retrieval work, it is necessary to rank them according to their relevance. This paper describes a relevance ranking method using information obtained from query word co-occurrences appearing in the retrieved documents. Distance between query words, their relative relationships in the database, and importance of query words are considered to decide the document relevance. Combined with traditional word frequency ranking, an overall relevance of retrieved documents is calculated. The traditional method alone and the combined method are compared using a test collection consisting of Japanese newspaper articles. Experimental results show that the proposed method improves retrieval recall about 0.098, or 37% compared to the traditional ranking method.

1. はじめに

新聞記事や特許などのテキスト情報が電子化されるようになり，大量の情報をデータベースに蓄積できるようになってきた．また，インターネット上にも大量のテキスト情報が存在している．大量のテキストから情報を取得する手段として，フルテキスト検索が一般的になっている．フルテキスト検索の利点として，検索語が文書中に出現しているものを漏れなく見つけ出せる点があげられる．その反面，検索語を含む文書数は膨大となることが多く，利用者が真に入手したい情報にたどりつくためには，絞り込み検索などの試行錯誤が必要である．そのため，利用者の労力を最小限に抑える検索システムの必要性が高まっており，近年，検索条件に対する文書の重要度を決定し，重要度順に検索結果を利用者に提示する手法が盛んに研究されている．米国では，1992 年より大規模なテキストデータを対象としたテキスト検索コンテスト TREC (Text REtrieval Conference) が行なわれており，1999 年には第 8 回目である TREC-8 が開催されている [13]．日本においても，日本語テキストの評価用テストコレクション BMIR-J1 や BMIR-J2 が構築されたり，TREC 同様コンテスト形式の NTCIR プロジェクトや IREX が実行されている [18),20),21),23)]．しかし，現状では TREC でも文書の重要度決定の精度は必ずしも高いとはいえず，精度の向上が望まれている．本論文は，重要度順に検索結果を提示するシステムにおける適合率の向上を目的とし，文書重要度の算出要素として，文書内での各検索語間の単語出現共起関係を用いた手法を検討する．本手法は，従来から広く利用されている検索単語の文書内出現頻度と出現文書頻度を使用した重要度付与 (TF-IDF) をベースと

† 株式会社ＮＴＴデータ 技術開発本部 オープンシステムセンタ
Open Systems Center, Research and Development Headquarters, NTT Data Corporation
†† 株式会社ＮＴＴデータ 技術開発本部 北米技術センタ
Technical Center of California, Research and Development Headquarters, NTT Data Corporation

している．まず，2章で基本的な重要度付与手法について述べ，3章で単語出現共起関係を用いた手法を説明する．4章で従来手法と提案手法を比較評価し，5章で評価結果の分析と考察を行う．6章では，今後の課題を述べる．

2. 従来の単語頻度情報による重要度付与

本章では，単語頻度情報を用いた従来の文書重要度付与方法について説明し，問題点をあげる．

2.1 単語共起出現による重要度付与

フルテキスト検索の場合，文書量の多いデータベースを検索したり，出現頻度の高い単語で検索すると，一般的にヒットする文書数が多くなる．このとき，利用者がすべてのヒット文書を参照して所望の文書を漏れなく探し出すことは現実的ではない．文書重要度付与は，利用者の検索要求に対して検索にヒットした文書に重要度を付与することであり，利用者は重要度の高い文書から参照することにより，素早く必要な文書を参照できる利点がある[8]．

重要度付与の方法として，単語の頻度情報を用いたアルゴリズムが広く検索システムで用いられている[7]．基本的なアルゴリズムは，単語の文書内出現頻度(TF: Term Frequency)，出現文書頻度(DF :Document Frequency)，および検索要求内出現頻度を用いて重要度を算出するものである．実際の重要度算出ではDFの逆数(IDF :$InverseDF$)を利用することからこの手法はTF-IDF法と呼ばれている．その他の統計情報として，検索対象の文書長を重要度算出に用いる場合が多く，最近のTRECでも主流になっている[10),13),14)]．

2.2 従来の文書重要度アルゴリズム

本節では文書重要度付与アルゴリズムとして，AT&TがTREC-6で採用したtf-idfアルゴリズムを説明する[12]．これは，Cornell大学が開発したSMARTシステムをベースとしたものである．本論文では，提案手法との比較のために本アルゴリズムを従来手法とした．AT&Tシステムのアルゴリズムを用いたのは検索精度が高いことがTRECで確認されている[2),13)]．SMARTシステムのアルゴリズムは，ベクトル空間モデルと呼ばれている[8]．ベクトル空間モデルでは，データベース内文書および検索要求内にユニークなS語の単語が出現した場合，データベース内の文書と検索要求のそれぞれをS次元の単語出現頻度として表現し，両者のベクトルの類似度により，文書の重要度を求めるものである．そのため，通常は各文書の単語出現頻度ベクトルをあらかじめ作成しておくことが多い

表1 パラメータの説明(1)
Table 1 Parameter Definition List(1)

パラメータ	説明
$word_i$	i 番目の検索語
D_k	検索対象の k 番目の文書
f_{tf}	文書内出現頻度による重要度算出関数
f_{idf}	出現文書頻度による重要度算出関数
f_{dlen}	文書長による重要度算出関数
$tf_i^{D_k}$	検索語 $word_i$ の文書 D_k 内出現頻度
df_i	検索語 $word_i$ の出現文書頻度
N	検索対象文書の総文書数
$length^{D_k}$	文書 D_k の文書長(文字数)
$avelen$	検索対象文書の平均文字数

が，本論文では検索要求に出現するM語の検索語のみのベクトルを生成する方法を取った．単語頻度情報を検索要求時に取得することにより，データベース内に存在する全単語の頻度情報を得る処理を省略した☆．次に，重要度の算出方法について説明する．式(1)のように検索要求QがM個の検索語で表わせる場合，文書D_kの重要度$score(D_k, Q)$は，前出した文書内出現頻度TF，出現文書頻度の逆数IDF，および重要度付与を行う文書の長さはそれぞれ各要素の関数f_{tf}，f_{idf}，f_{dlen}により表せる．

$$Q = (word_1, \cdots, word_i, \cdots, word_M) \quad (1)$$

$$score(D_k, Q) = \sum_{i=1}^{M} \left[f_{tf}(tf_i^{D_k}) \cdot f_{idf}(df_i) \cdot f_{dlen}(D_k) \right] \quad (2)$$

ただし，

$$f_{tf}(tf_i^{D_k}) = tf_i^{D_k} \quad (3)$$

$$f_{idf}(df_i) = \log\left(\frac{N+1}{df_i}\right) \quad (4)$$

$$f_{dlen}(D_k) = \frac{1}{0.8 + 0.2 \times \dfrac{length^{D_k}}{avelen}} \quad (5)$$

ここで，

$tf_i^{D_k}$は文書D_k内に$word_i$が出現する頻度，

df_iは検索対象のデータベース内で$word_i$が出現する文書の頻度（文書数），

Nは検索対象のデータベースの総文書数，

$length^{D_k}$は文書D_kの文字数，

$avelen$は検索対象文書の平均文字数である．

式(5)では，Pivoted Document Length Normaliza-

☆ 日本語テキストを対象とする場合，後述する形態素解析の問題があることや，データベース内の文書が更新されたときに，全単語の頻度情報を取得する処理が必要となるため，検索時に単語の頻度を取得することとした．

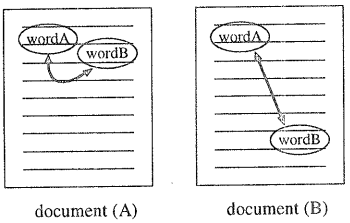

図1 検索語出現位置と重要度の関係
Fig. 1 Relationship between distance of query words and document relevance

tion[11] を用いて文書の長さによる正規化処理が行なわれている．文書の長さによる正規化処理として文書のバイト数が用いており，本論文でも文書の文字数による正規化を行うこととする[12]．

2.3 従来の文書重要度付与の問題点

まず，従来手法である TF-IDF 法の問題点を述べる．TF-IDF 法では，検索語の文書内出現頻度と文書出現頻度による重要度算出を行っており，検索語の出現頻度が同じ場合，文書の重要度は等しいものとする．しかし，検索語頻度が同数の場合でも，検索語の出現位置により文書の重要度が異なることも考えられる．この場合，検索語の出現頻度情報だけでは文書重要度を変えることができない．たとえば，図1に検索語の出現箇所が異なる文書を示す．(A) の文書では，複数の検索語同士が近接して出現しており，2つの検索語の関連が高い場合，その出現部分については局所的に重要であり，文書の主題となる文を含んでいる可能性が高い．一方，(B) の文書では，検索語が遠くに離れて出現しており，2つの検索語の関連性は (A) に比べ低いと考えられる．このことから，検索語の出現位置を重要度付与の要素にいれることが有効であると考えられる．

検索語の出現位置を用いた手法としては，文書の構造情報（タイトル，概要，本文など）を利用したものや[15]，検索語の共起出現情報を利用したものがある．構造情報の利用では，検索語が出現する項目によって重要度を変えることで文書全体の重要度を算出する方法を取っている．しかし，この手法の場合，文書の種類により項目が異なるため，重要度を文書の種類に対応して変える必要がある．

共起出現情報を用いた手法として，共起検索語同士の文書中での出現距離を用いたものがある[4]．この手法は，共起を考慮する複数の検索語が文書内に出現する場合，共起している検索語の間に含まれる非検索語の単語数によって出現距離を測定し，その距離が近い文書には高い重要度を付与している．具体的には，検索語 $word_A$ と検索語 $word_B$ が文書中に共起出現している場合，$word_A$ $word_X$ $word_Y$ $word_B$ の順に出現している場合と $word_A$ $word_X$ $word_B$ の順に出現している場合では，それぞれ共起検索語の間には2語，1語あるため，非検索語の数が少ない後者に高い重要度を付与するものである．この手法を用いた英語テキストでの評価実験により，単語頻度のみを用いた重要度付与方法に比べ適合率が0.1程度向上することを示している[4]．しかし，この手法は検索語 $word_A$ と検索語 $word_B$ の関連性は考慮していないため，検索語 $word_A$ と検索語 $word_B$ の関連性が低い単語であったり，重要でない単語と共起している場合にも文書の重要度を高めることになり，検索精度が低下することがあった．本論文では，共起検索語の出現距離だけでなく，共起検索語のそれぞれの単語の重要度や検索語同士の関係の共起出現重要度を考慮した手法を提案する．

3. 単語共起出現による重要度付与方法

上述の単語頻度による重要度付与アルゴリズムでは，単語の文書内頻度 (TF) と文書頻度 (DF) を基本特徴量として用いるが，高精度な重要度付与を目的とした新たな重要度算出のための特徴量を提案する．また，検索語の共起に関する特徴を文書の重要度付与に反映させる方法を説明する．なお，本検討では，検索要求は自然語文で与えるのではなく，形態素解析によって単語に分割されていることを前提とする．

3.1 単語共起出現の特徴量

本論文では，共起出現における検索語同士の関係も考慮した文書の重要度付与手法を検討する．本手法は，文書内に出現した検索語同士の関係を，近接出現距離，共起関連度，共起検索語の重要度という3つの特徴量で表わし，これらの要素により該当する検索語に対する共起出現による重要度を決定するものである．

具体的には，文書中に出現した任意の検索語 $word_A$ に着目して次のことを考える．通常，$word_A$ の文書内出現頻度値 (TF) はある文書中に $word_A$ が出現した数である．この $word_A$ に関して他の検索語 $word_B$ が近接して出現し，単語共起関係がある場合，当該の $word_A$ に対する出現頻度値に単語共起関係についての重要度を加えて文書の重要度を高めることを行う．すなわち，同じ検索語 $word_A$ であっても，近接する他の検索語により $word_A$ の重要性が異なるようにする．なお，タイトル，概要などの文書の構造情報の利用は本論文では検討しないこととする．また，共起関係を用いることから，複数の検索語が指定された場合の検索への適用を考える．次に，検索語の共起出現に

関連する3つの特徴量について説明する．

3.1.1 近接出現距離の利用

関連のある検索語同士が，文書内のある部分にまとまって出現した場合，この部分は検索要求を満している部分であると考えられる．また，このような部分を含む文書と含まない文書では，前者の方が主題を的確に表わしているため文書全体として重要度も高いと考えられる．このことより，文書の重要度に関して次の仮説を立てる．

- 近接出現距離に関する仮説：
  関連のある検索語が一文書内で近接して出現する文書は検索語との適合度が高い

この仮説は，局所的に複数の検索語のAND検索条件が成り立っていると考えることができる．OR検索の結果にはAND検索の結果を含んでおり，重要度を付与し文書を重要度順に提示する場合，一般にAND検索条件に適合する文書は，OR検索条件に適合する文書より重要度が高くなる傾向があるので，本仮説は有効であると考えられる．

3.1.2 検索語間の共起関連性の利用

検索語の中には互いに関連がない単語も含まれているため，「関連のある検索語」を決定する必要がある．ここで，検索語同士の関連性を関連度という指標で表わす．この関連度について以下の仮説を立てる．

- 検索語間の共起関連性に関する仮説：
  近接出現する検索語同士の関連度により共起の重要性は変わる．重要度は検索語同士が近接して共起する確率（共起係数）とする．

「関連ある検索語」とは，検索語同士が主従関係や修飾関係にある場合や，複合語を構成する語となる場合であり，文書中で近接して出現する確率が高いと考えられる．そのため，この確率を表わす共起係数を検索語の関連度と考える．これにより，共起係数が高いものほど，検索語同士の関連度は高いと考えることができる．なお，共起係数は，文書データベース中の統計情報から算出することができる．

3.1.3 共起検索語の重要度の利用

共起出現している二つの検索語のうち，着目している検索語に対するもう一つの検索語を共起検索語と呼ぶ．ここで，検索語 $word_A$ と共起検索語 $word_B$ が文書中に出現しているときを考える． $word_A$ に着目した場合，前節で示した共起係数が同じ場合でも，共起検索語 $word_B$ が検索要求に対して重要な単語である場合と，そうでない場合が考えられる．重要でない共起検索語にもかかわらず，近接出現による重要度を付加することは意味がない．共起検索語の重要度を考慮するため，次の仮説を立てる．

- 共起検索語の重要性に関する仮説：
  共起検索語の重要度が高い場合，共起出現に対する重要度も高い．

検索語同士が近接して共起出現することによる文書重要度（以下，共起重要度と呼ぶ）を最終的な文書重要度に反映させる方法，および近接の定義については，次節で述べる．

3.2 共起出現の定義方法

検索語 $word_A$ と検索語 $word_B$ の「共起出現」を $word_A$ と $word_B$ の出現間隔が距離 $dist$ 以内であるとき，「$word_A$ と $word_B$ は共起出現する」と定義する．

共起出現には様々な定義があり，注目する複数の単語が，どの単位（たとえば，段落単位や文単位など）で出現するかを考慮することができる．ここでは，共起出現を相互の検索語が一定の距離 $dist$ 以内に現われた場合とする．距離 $dist$ を文書，段落，文の数と文字数で定義する．近傍の定義として，単語数が用いられているものが多いが[3),16)]，本論文では文字数を用いた．その理由は，英語等のテキストでは，スペース文字により各単語の区切りが明確であるが，日本語テキストを対象とする場合，形態素解析等により単語切り出し（単語分割）を行う必要があり，

(1) 形態素解析の単語分割誤りの発生，
(2) 複合語の単語分割基準のあいまいさ，
(3) 処理速度の低下

などの問題が存在するためである．

3.3 共起重要度の算出方法

3.1節で提案した，検索語の近接共起出現を文書重要度に反映させる方法について述べる．上記基本アルゴリズムでは，算出要素は $tf_i^{D_k}$ と df_i であるが，共起出現によって与えられる共起重要度 cw を $tf_i^{D_k}$ に加算することで文書重要度に反映させる．

ここで，文書 D_k 内の $word_i$ の出現集合を $G_i^{D_k}$，文書 D_k 内の $word_j$ $(i \neq j)$ の出現集合を $G_j^{D_k}$ としたとき，新たに $tf_k^{D_i\prime}$ を文書内出現頻度として重要度を算出する．$tf_k^{D_i\prime}$ は次式により算出する．

$$tf_i^{D_k\prime} = tf_i^{D_k} + \sum_{a \in G_i^{D_k}} \sum_{b \in G_j^{D_k}} cw_{ab}^{D_k} \qquad (6)$$

ここで，右辺第2項は，文書 D_i 内に出現している $word_i$ と $word_j$ $(i \neq j)$ の検索語との共起重要度の総和を示している．

次に共起重要度 cw の算出方法について説明する．ここで，3.1節で提案した各共起情報を数値化して係数として定義し，各係数を用いて共起重要度 cw を算

表 2 パラメータの説明 (2)
Table 2 Parameter Definition List(2)

パラメータ	説明
$\rho(word_A, word_B)$	$word_A$ と $word_B$ における検索語間の近接出現距離係数
$\sigma(word_A, word_B)$	$word_A$ と $word_B$ における検索語間の共起係数
$\tau(word_A, word_B)$	$word_A$ の共起語 $word_B$ に対する近接出現共起単語の重要度係数
d	近接出現距離の設定しきい値
$dist_{AB}$	$word_A$ と $word_B$ 出現距離
atf_A	データベース内の $word_A$ 出現頻度総数
rtf_{AB}	近接出現距離 d 以内に $word_B$ が現われる $word_A$ の出現頻度
δ	共起重要度正規化係数

出する.

- $\rho(word_A, word_B)$: $word_A$ と $word_B$ における検索語間の近接出現距離係数
- $\sigma(word_A, word_B)$: $word_A$ と $word_B$ における検索語間の共起係数
- $\tau(word_A, word_B)$: $word_A$ の共起語 $word_B$ に対する近接出現共起単語の重要度係数

まず, $\rho(word_A, word_B)$ と $\sigma(word_A, word_B)$ を次のように定義する.

$$\rho(word_A, word_B) = \begin{cases} \dfrac{(d+1) - dist_{AB}}{d+1} & : if\ d \geq dist_{AB} \\ 0 & : otherwise \end{cases} \quad (7)$$

$$\sigma(word_A, word_B) = \frac{rtf_{AB}}{atf_A} \quad (8)$$

ただし,

d は近接出現距離の設定しきい値,

$dist_{AB}$ は $word_A$ と $word_B$ の出現距離,

atf_A はデータベース内の $word_A$ 出現頻度総数,

rtf_{AB} は近接出現距離 d 以内に $word_B$ が現われる $word_A$ の数.

ここで, $\rho(word_A, word_B)$ は 0 から 1 の値を取り, 近接して出現しているほど, $\rho(word_A, word_B)$ は 1 に近い値を取る. また, 共起係数 $\sigma(word_A, word_B)$ は, 検索対象のデータベースから取得することにより, データベースに依存した共起係数を得ることができる.

$\sigma(word_A, word_B)$ も 0 から 1 の値を取り, 常に近接して出現する単語同士の場合, $\sigma(word_A, word_B)$ の値は大きくなる.

$\tau(word_A, word_B)$ を次のように定義する.

$$\tau(word_A, word_B) = f_{idf}(df_B) = \log(\frac{N+1}{df_B}) \quad (9)$$

ただし, N は検索対象のデータベースの総文書数である. これは, 近接出現共起単語の重要度を, 文書出現頻度を利用して求めるものである. 算出方法は式 (4) と同様である. $word_A$ に着目した場合, $word_A$ 以外の重要度を考慮するため, 式 (9) では $word_A$ の重要度は無関係である.

最後に, 近接出現係数 $\rho(word_A, word_B)$, 共起係数 $\sigma(word_A, word_B)$, 近接出現共起単語重要度 $\tau(word_A, word_B)$ と共起重要度 cw_{AB} の関係を示す. これまで述べてきたように, $\rho(word_A, word_B)$, $\sigma(word_A, word_B)$, $\tau(word_A, word_B)$ は共起重要度 cw に対して互いに独立した関係ではなく, 他の係数との相互作用があると考え, $word_A$ と $word_B$ と共起重要度 cw_{AB} を次の式で表わすこととした.

$$cw_{AB} = \rho(word_A, word_B) \times \sigma(word_A, word_B) \times \tau(word_A, word_B) \times \delta \quad (10)$$

ここで, δ は共起重要度正規化係数である.

各係数の積で共起重要度 cw_{AB} を定義したしたのは, 3 つの共起重要度係数のうち, いずれかの係数の値が小さい場合, 共起重要度の値も小さく設定するようにしたためである.

4. 評 価

本論文で提案した単語共起出現関係を用いた文書重要度の付与手法について, 検索精度を評価する.

4.1 評価条件

本節では, 評価条件について説明する.

4.1.1 評価対象データベース

評価には, 日本語の情報検索評価用テストコレクションである, BMIR-J2 ☆ を使用した [18].

BMIR-J2 は, 新聞記事 5080 件, 検索要求文 60 件, および各検索要求文に対する正解集合から構成されている. 正解集合には 3 つのランクが付与されており, その基準は表 3 に示す通りである. 本評価では A ランクと B ランクを正解とした. また, BMIR-J2 の基本セットである 50 件の検索要求を用いた. ☆☆

☆ (社) 情報処理学会・データベースシステム研究会が, 新情報処理開発機構との共同作業により, 毎日新聞 CD-ROM'94 データ版を基に構築した情報検索システム評価用テストコレクション

☆☆ BMIR-J2 では, 60 件の検索要求を「A 判定, B 判定の正解文書数」≧ 5 を満たす標準セット 50 件とそれ以外の追加セット 10 件に層別しており, 検索精度の評価には基本セットの使用を推奨している. [22]

表 3 BMIR-J2 の正解基準
Table 3 Rank of relevancy assessment for BMIR-J2

ランク	基準
A	検索要求を主題とする記事
B	記事の主題は検索要求と異なるが，検索要求の内容を少しでも記述している記事
C	全く関連のない記事（不正解）

表 4 評価で用いたパラメータ
Table 4 Parameters used in the experiment

パラメータ		値
近接出現距離の設定しきい値	文字数 d_{char}	10,20,30,50,70,100
	文数 d_{sen}	0,1,2,3,5,7,10
	段落数 d_{para}	0,1,2,3,5,7,10
正規化係数	δ	0,0.1,0.2,0.3,0.5,0.7, 1.0,2.0,3.0,5.0,7.0, 10.0,20.0,30.0,50.0,70.0

4.1.2 評価方法

各検索要求文から，次節で示す方法で検索語を抽出し検索処理を行い，重要度順に出力された検索結果に対し評価を実施する．なお，重要度付与に用いた文書領域は見出しおよび本文部分に限定した．評価指標として，適合率 (Precision)，再現率 (Recall) を用いた．適合率は，正しい文書のみを検索している度合い，再現率は，正しい文書を漏れなく検索している度合いを示しており，次の式で計算される．

$$\text{適合率} = \frac{\text{検索された正解文書数}}{\text{検索されたすべての文書数}} \quad (11)$$

$$\text{再現率} = \frac{\text{検索された正解文書数}}{\text{すべての正解文書数}} \quad (12)$$

各検索要求ごとに，再現率が 0.0,0.1, …,1.0 の場合☆の適合率を算出し，各検索要求における適合率の平均値である平均適合率を算出し評価指標とした[7),8)]．また，別の評価指標として，改善率を定義する．各検索要求に対して，完全な重要度付与ができた場合，平均適合率は 1.0 となるが，比較対象とする重要度付与手法に対して，どの程度完全な重要度付与結果に近付いたかを示すものである．改善率は次の式で計算する．

$$\text{改善率} = \frac{avepre_B - avepre_A}{1.0 - avepre_A} \quad (13)$$

ただし，

$avepre_A$ は従来手法による平均適合率

$avepre_B$ は提案手法による平均適合率

ここで，評価では，

(1) 単語頻度情報のみによる検索精度（従来手法）

☆ 再現率が低い (0 に近い) 場合は，重要度付与の結果，ランク上位の部分を示し，再現率が高い (1 に近い) 場合は，ランク下位の部分を示す．

表 5 抽出された検索語の例
Table 5 Examples of extracted query words

検索要求文	抽出検索語
・菓子メーカー	菓子，メーカー
・半導体製品の生産	半導体，製品，生産
・業績悪化を原因とする企業合併の事例	業績，悪化，原因，企業，合併，事例

表 6 検索要求と検索対象文書の規模
Table 6 Statistics of queries and retrieved documents

	BMIR-J2 検索対象文書	評価対象 検索要求文	評価対象 検索語の組
件数	5080	44	44
最大	999 文字	18 文字	7 語
最小	400 文字	3 文字	2 語
平均	618 文字	9.4 文字	3.5 語
	11.9 文		
	5.3 段落		

(2) 単語頻度情報と共起重要度による検索精度（提案手法）

の両者を比較することにより，共起重要度の効果を評価する．共起出現の定義を，同一の文書内に $word_A$ と $word_B$ が共に出現し，かつ両者の単語が近接出現距離の設定しきい値 d 以内に出現している場合とする．ここで，近接出現距離の定義を，

(1) 文字数による計測

(2) 文数による計測

(3) 段落数による計測

(4) 同一文書内に出現した単語はすべて共起出現している単語とする

として，実験を行った☆☆．近接出現距離の設定しきい値 d と，正規化係数 δ を表 4 に示すように変化させ，検索精度の測定を行う．ここで，正規化係数 δ は共起重要度の影響の大きさを変化させるものである．

また，本手法で用いる共起重要度の 3 つの各係数をすべて利用するのではなく，1 ～ 2 個の要素を用いて共起重要度を算出した場合の検索精度の変化を調べる．ここで，共起重要度の算出要素として用いない係数は，式 (10) の該当する項を 1 とする．

4.1.3 検索語の生成

BMIR-J2 では，検索要求は単語ではなく表 5 の左列に示すようなフレーズの形で提供されている．本システムの前提となっている最適照合法 (best match) では，自然言語で与えられた検索要求に対する各文書のスコアを付与するものである．本システムの前提条

☆☆ 文書数，文数による計測では，それぞれ同一文，同一段落以外に出現している共起検索語も共起出現としている．

表 7 平均適合率 (近接出現の定義：文字数)
Table 7 Mean precision (co-occurrence measures: number of characters)

		正規化係数 δ											
d_{char}	0.0	0.1	0.2	0.3	0.5	1.0	2.0	3.0	5.0	10.0	20.0	30.0	50.0
10		0.265	0.267	0.269	0.272	0.280	0.291	0.297	0.306	0.320	0.328	0.330	0.331
20		0.265	0.269	0.273	0.276	0.288	0.299	0.308	0.318	0.331	0.339	0.339	0.339
30	0.262	0.267	0.272	0.275	0.280	0.292	0.306	0.316	0.327	0.339	0.347	0.350	0.349
50		0.268	0.273	0.279	0.286	0.302	0.317	0.326	0.336	0.345	0.349	**0.351**	0.349
100		0.271	0.279	0.285	0.293	0.309	0.324	0.330	0.340	0.346	0.347	0.347	0.348

表 8 平均適合率 (近接出現の定義：文数)
Table 8 Mean precision (co-occurrence measures: number of sentences)

		正規化係数 δ											
d_{sent}	0.0	0.1	0.2	0.3	0.5	1.0	2.0	3.0	5.0	10.0	20.0	30.0	50.0
0		0.270	0.279	0.283	0.293	0.307	0.319	0.325	0.336	0.341	0.345	0.347	0.347
1		0.280	0.292	0.301	0.311	0.322	0.337	0.343	0.348	0.351	0.351	0.351	0.351
2	0.262	0.288	0.301	0.311	0.319	0.329	0.341	0.345	0.349	0.351	0.351	0.351	0.350
3		0.295	0.309	0.316	0.323	0.338	0.348	0.351	0.354	0.355	0.354	0.355	0.356
5		0.306	0.316	0.322	0.333	0.346	0.352	0.353	0.354	0.355	0.357	**0.358**	**0.358**
10		0.311	0.322	0.332	0.340	0.349	0.351	0.352	0.353	0.354	0.355	0.355	0.355

表 9 平均適合率 (近接出現の定義：段落数)
Table 9 Mean precision (co-occurrence measures: number of paragraph)

		正規化係数 δ											
d_{para}	0.0	0.1	0.2	0.3	0.5	1.0	2.0	3.0	5.0	10.0	20.0	30.0	50.0
0		0.281	0.294	0.300	0.310	0.322	0.334	0.338	0.342	0.345	0.344	0.344	0.343
1		0.299	0.313	0.320	0.330	0.343	0.352	0.353	0.355	0.355	0.356	0.357	0.357
2	0.262	0.308	0.319	0.325	0.338	0.350	0.357	0.357	0.358	0.358	**0.360**	0.358	0.358
3		0.312	0.322	0.332	0.339	0.348	0.354	0.354	0.354	0.356	0.357	0.357	0.357
5		0.313	0.326	0.335	0.341	0.349	0.351	0.352	0.354	0.354	0.353	0.353	0.353
10		0.316	0.326	0.335	0.341	0.349	0.351	0.352	0.353	0.353	0.353	0.353	0.353

件は検索要求を検索語に分割して与えるものであるため，検索要求文から検索語を生成する必要がある．

本評価では，検索要求文から「茶筌」[25] を用いて形態素解析を行い，抽出された形態素の中から検索語を選択した☆．表 5 の右列に検索語抽出例を示す．本提案手法では，文書内の検索語の共起を用いて重要度付与を行うものであるため，検索語が 1 個であるものには，重要度の変化は起こらない．そのため，従来手法との比較では，検索語が 2 語以上の検索要求だけを評価対象とする．

以下の評価では，44 件の検索要求についての検索結果を対象とする．なお，評価に用いた BMIR-J2 の検索要求や検索対象文書の大きさは表 6 に示すとおりである．

☆ 検索語とした品詞：普通名詞，名詞性名詞助数辞，固有名詞，サ変名詞，形容詞，名詞性名詞接尾辞，ナ形容詞接頭辞，未定義語
検索語としなかった品詞：数詞，名詞接続助詞，格助詞，動詞，記号，副詞的名詞

5. 評価結果の分析と考察

本章では，実験から得られたデータを分析し，共起重要度を利用した場合の効果を考察する．

5.1 評価結果

まず，検索精度が近接出現距離の定義方法により，どのように変化するかを評価した．近接出現距離を，文字数，文数，段落数により定義した場合の平均適合率をそれぞれ，表 7，表 8，表 9 に示す．

また，近接出現距離の各定義での最も平均適合率が高かったものを表 10 に示す．ここで，従来手法の平均適合率は 0.262 である．なお，重要度付与した文書（いずれかの検索語を含む記事）の数は平均 1222 文書，この中で A または B ランクの正解が付与されているものは平均 26.7 文書，一検索要求あたりの本システムの正解出力数は平均 23.6 文書であった．このときの適合率は 0.019，再現率は 0.884 である．フルテキスト検索を行った場合，このように適合率が非常に低下するため，重要度付与の必要性は明らかである．

今回評価を行った近接出現距離のすべての定義方法

表 10 近接出現距離の定義による適合率 (各定義での最適適合率)
Table 10 Precision by distance measures of co-occurrence

近接出現距離の定義方法	近接出現距離のしきい値	正規化係数 δ	適合率
文字数	$d_{char} = 50$	30.0	0.351
文数	$d_{sent} = 5$	30.0	0.358
段落数	$d_{para} = 2$	20.0	0.360
文書内		3.0	0.349

表 11 共起重要度の算出要素の組み合わせによる適合率 (要素組合せでの最適適合率,$d_{para} = 2$ のとき)
Table 11 Precision by element combinations of co-occurrence importance ($d_{para} = 2$)

共起重要度 共起距離係数 ρ	共起係数 σ	共起単語係数 τ	正規化係数 δ	適合率
○	×	×	1.0	0.316
×	○	×	30.0	0.337
×	×	○	0.5	0.336
○	○	×	50.0	0.344
○	×	○	1.0	0.349
×	○	○	20.0	0.345
○	○	○	20.0	0.360

○：共起重要度算出に用いた要素
×：共起重要度算出に用いない要素

表 12 適合率 ($d_{para} = 2, \delta = 20.0$ のとき)
Table 12 Precision($d_{para} = 2, \delta = 20.0$)

再現率	従来手法	提案手法
0.00	0.628	0.680
0.10	0.458	0.576
0.20	0.387	0.506
0.30	0.336	0.460
0.40	0.295	0.423
0.50	0.255	0.363
0.60	0.219	0.316
0.70	0.178	0.264
0.80	0.158	0.237
0.90	0.116	0.161
1.00	0.094	0.130
平均	0.262	0.360

表 13 適合率 ($d_{para} = 2, \delta = 20.0$ のとき)
Table 13 Precision ($d_{para} = 2, \delta = 20.0$)

再現率	従来手法	提案手法
上位 5 文書	0.373	0.491
上位 10 文書	0.323	0.416
上位 15 文書	0.292	0.371
上位 20 文書	0.276	0.349
上位 30 文書	0.243	0.289
平均	0.265	0.352

において，平均適合率の向上がみられた．すべてのパラメータにおける実験結果の分析をここで示すことができないので，近接出現距離を段落数により定義した $d_{para} = 2$，$\delta = 20.0$ の場合を中心に評価結果の分析と考察を行う．また，$d_{para} = 2$，$\delta = 20.0$ のとき共起重要度算出要素を 1 ～ 2 個組み合わせた場合の，平均適合率が最も高い場合を表 11 に示す．

5.2 提案手法の有効性の分析

$d_{para} = 2$，$\delta = 20.0$ の場合，提案手法により平均適合率は 0.098 向上 (従来手法と比較した場合の向上率は 37.4%) が見られた (表 12)．また，表 13 は重要度付与された検索結果の上位 n 件の適合率を従来手法と比較したものである．一般的に検索システムの利用者は，重要度順に出力されたすべての結果を閲覧することは現実的ではないため，上位 n 件での適合率の差が実用的な評価指標として用いられることが多い．上位 20 位での適合率の差は 0.073 であり提案手法が正解を上位に位置付けていることがわかる．

BMIR-J1 を用いたシステムの検索精度の比較は符合検定により行うことが推奨されている [20),22)]．本評価で用いた BMIR-J2 でも，符合検定により従来手法と提案手法の比較を行う [26)]．帰無仮説 H_0「従来方法 X_1 と提案方法 X_2 の平均適合率に差はない」，対立仮説 H_1「従来方法 X_1 と提案方法 X_2 の平均適合率に差がある」としたとき，X_1 と X_2 による各検索要求 Q_j に対する平均適合率をそれぞれ $avepre_{Q_j}^{X_1}$，$avepre_{Q_j}^{X_2}$ とする．このとき，$avepre_{Q_j}^{X_1} > avepre_{Q_j}^{X_2}$ の数を n_p，$avepre_{Q_j}^{X_1} < avepre_{Q_j}^{X_2}$ の数を n_m とする．帰無仮説 H_0 のもとでは，$n_p = n_m$ となる．ここで，$d_{para} = 2$，$\delta = 20.0$ の場合，評価対象とした 44 個の検索要求の中で，平均適合率の向上があったものは 32 検索要求，低下したものは 8 検索要求，変化のなかったものは 4 検索要求であったので，$n_p = 32$，$n_m = 8$ である．信頼区間 α に対して，

$$P_r(X \geq X_L) \leq \alpha \tag{14}$$

を満たす X_L は，$n_p + n_m$ が 30 以上なので，

$$X_L = \frac{(n_p + n_m) + 1 + (n_p + n_m)^{\frac{1}{2}} z(\alpha)}{2} \tag{15}$$

で算出できる．信頼区間を 99%としたとき，$z(\alpha) = 2.57$ であるから，$X_L = 28.63$ となる．ここで，$n_p = 32 > 28.63$ であるから，帰無仮説 H_0 は棄却され，対立仮説 H_1「従来方法 X_1 と提案方法 X_2 の平均適合率に差がある」が信頼区間を 99%で採用され，本提案手法の効果が符合検定により示される．

また，従来手法と提案手法の平均適合率の差について検定を行うと，平均適合率の差が 0.06 のとき，信頼区間 95%で有意な差がある．すなわち，提案手法は従来手法と比較して，適合率の差が少なくとも 0.06 (割合にして 23%の向上) 以上あることが示される．ま

表 14 検索精度 (改善率) の変化が大きい検索要求
Table 14 Topics with big change in precision

番号	検索要求	改善率	適合率の向上
108	携帯電話またはパーソナルハンディホン	0.868	0.171
110	衛星放送	0.767	0.391
126	地価の下落	0.730	0.360
136	赤字国債の発行	0.658	0.156
120	所得税の減税	0.627	0.133
122	政党に対する献金	0.552	0.348
125	マンションの販売	0.550	0.329
142	材料・設備の現地調達	0.529	0.456
116	3期以上連続の減益企業	0.404	0.162
123	国連軍派遣	0.378	0.213
114	教育産業	-0.093	-0.083
131	株価動向	-0.076	-0.042
118	半導体製品の生産	-0.067	-0.054
124	電気通信に関する規制緩和	-0.045	-0.038
148	業績不振の責任を取った経営者	-0.033	-0.031

た，表 14 に改善率の向上や低下が大きい検索要求を示す．改善率の向上が見られた検索要求は向上度合が大きく，逆に改善率が低下した検索要求は，低下度合はそれほど大きくないことがわかる．改善率の低下が大きい検索要求について分析すると，検索語が共起出現していない文書が正解文書である場合が多い．これは，検索語が共起出現している不正解文書の重要度を高めたことにより適合率の低下が起きたものである．しかし，該当する共起検索語間の共起重要度も小さいため，適合率の低下は小さくなっていると考えられる．

5.3 共起検索語の出現

文書中に複数の検索語が共起するのは，

- 複合語を構成するいくつかの基本語が文書中に共起して出現した
- 主題が重点的に記述されている箇所が存在しているため検索語が共起して出現した

ことが考えられる．改善率の向上が最も大きかった検索要求文に含まれる「携帯電話」は，「携帯」と「電話」の 2 語が検索語として抽出された．「携帯電話」として複合語を検索語した場合，検索が不可能な「....携帯・自動車電話....」という部分は，基本語を検索語とすることで，検索が可能となった．本評価で用いた形態素解析は「茶筌」であるが，他の形態素解析器「Majesty」[19] と比較すると，複合語としてではなく，基本語が形態素として出力される．検索語を基本語単位とすることで，複合語が検索語として与えられた場合に文字列がマッチングしないために起こる検索漏れを回避することができるとともに，本提案手法で用いる検索語同士の近接出現情報を用いることにより，単語と単語の間に助詞や関係句を含んだ文書であっても的確な検索結果を出力することが可能となる．検索語を複合語としたときの平均適合率は 0.231，一検索要求あたりの正解出力数は平均 20.0 文書であった．基本語を用いた場合の平均適合率 0.262 と平均出力正解数 23.6 と比較した場合値の低下がみられ，検索漏れの発生が多く観察されたため，検索語として基本語を用いることは有効であると考える．また，一般的に文書には複数の話題を含むものが多いが，一般的な文書の重要度付与は，文書長の要素も算出要素としているため，異なる話題を多く含むような長い文書に対する重要度は相対的に低くなる傾向がある．本手法を用いることで，検索要求に適合した話題を含むものの重要度を適切に上げることが可能となる．

5.4 適用共起重要度要素

本手法では，共起重要度の算出要素として，近接出現係数 ρ，共起係数 σ，共起単語重要度係数 τ を用いた．各係数要素および，共起距離のしきい値 d と正規化係数 δ について考察する．

- 平均適合率の向上度合いは，共起重要度の算出要素を単独あるいは 2 つ用いた場合と算出要素の 3 つをすべて使用した場合を比較したすると，後者の効果が最も効果が高い (表 11)．本算出要素は，組合せにより効果があるといえる．
- 検索対象の文書中に出現した各検索語について各共起重要度係数値 ρ,σ,τ の相互および共起重要度 cw との相関を分析すると，各相関値ともに 0.05 以下であり，相関は低い．それぞれ独立した特徴値であることがわかる．
- 本評価では，近接出現距離の設定しきい値を変化させて評価を行ったところ，最も精度向上がみられたのは，それぞれ，$d_{char} = 50$，$d_{sent} = 5$，$d_{para} = 2$ のときであった．日本語の新聞記事を検索対象とした場合，検索対象文書の文字数，文数，段落数の平均値のそれぞれ 1/10,2/5,2/5 程度とすることが望ましいと考えられる．ただし，他の種類の文書に適用する場合は，文や段落といった文書構造が異なるために本評価で得られたしきい値を用いることはできない．また，文字数に関するしきい値は，言語に依存するものであり，多くの日本語の文書に対しては本評価結果である $d_{char} = 50$ を適用することが望ましいと考える．
- 単純に文字数で共起単語の定義をするよりも，意味のある文字列の塊である，文や段落を共起出現距離とした方が，若干効果が高いことがわかる (表 10)．評価に用いた BMIR-J2 内の文書は，文や段落等の文書構造の区切りが明記されているが，

表 15 処理時間
Table 15 Response time

処理時間	従来手法 (sec)	提案手法 (sec)	従来手法に対する比率	ヒット文書数
平均	2.30	64.46	23.13	1196
最小	0.63	1.23	1.95	14
最大	6.52	284.10	45.35	4226

一般的な文書は，文書構造が明確でないものが多い．その場合，近接共起距離の定義を文字数にしても，精度の向上は十分期待できる．

- 式 (6) の右辺第 2 項は，単語出現頻度に加算する共起重要度であるが，本共起重要度の大きさは，おおむね右辺第 1 項と比較しても十分大きい値であった．共起重要度は十分に重要度付与に寄与している．

5.5 処理時間

利用者に対して対話型の処理を行う検索システムの場合，検索精度と同様に高速な検索速度も期待される．提案手法では各検索語の共起情報をデータベースから取得しているために，特に共起出現距離の取得において，検索語の出現頻度が高く場合，検索処理に時間を要する．表 15 に従来手法と提案手法の処理時間の比較を示す．本評価では，Sun Ultra2(CPU:UltraSPARC2 200MHz, メモリ 512MB) を用いて測定した．測定結果を分析すると，共起情報の取得件数に影響をおよぼすヒット件数の増加に伴い処理時間の増加がみられた．検索対象文書数が大きいデータベースの検索においては，従来手法により初期検索を行い，文書重要度が高い文書に対してのみ本手法の適用を選択的に行うことにより処理時間の増加を抑えることが可能であると考える．BMIR-J2 による評価では，従来手法を初期検索とし，その結果の上位 200 文書に対して提案手法を適用したところ，$d_{para} = 2, \delta = 20.0$ のとき平均適合率は 0.360 となり，全文書に適用した場合と同程度の平均適合率を得ることができた．また，処理速度も従来手法の 6.17 倍に抑えることが可能となった．

6. おわりに

本論文では，文書重要度付与における精度向上のため，検索語の共起重要度を用いた手法を検討した．本手法では，複数の検索語間の文書内共起出現情報として，近接出現距離，共起関連度，共起検索語の重要度という特徴量を用いて共起出現重要度を決定した．また，単語頻度情報を用いた重要度付与手法と組み合わせることにより，平均適合率が約 0.098 向上 (従来手法と比較した場合の向上率は 37%) することを，日本語の情報検索評価用テストコレクション (BMIR-J2) によって示した．適合率の向上のためには，検索語の自動展開により関連語を検索語に加えることも必要であり，その効果も確認されている [14]．展開された検索語の共起情報を用いることにより，再現率も含めてさらに高い検索精度を得ることが期待できる．一方，検索システムの観点から見た場合，検索精度と同様に検索速度の向上も期待されるため，検索精度と検索速度のバランスのとれた検索手法の検討も重要である．今回評価では BMIR-J2 を用いたが，実際の検索システムでは，検索対象の文書も多様であることから，対象文書に最適なパラメータ設定方法や今回検討した近接出現係数や共起係数の実装方法についても今後さらに検討を進めていく．さらに，同じ検索語であっても利用者によって検索意図が異なることがある．個々の利用者の検索意図を把握した上で重要度付与を行うことも検討をしていく．

参考文献

1) C. Buckley: The importance of proper weighting methods, In M. Bates, editor, Human Language Thchnology, Morgan Kaufman, 1993.
2) C. Buckley, G. Salton, J. Allan and A. Singhal: Automatic Query Expansion Using SMART: TREC3, Proc. of The 3rd Text Retrieval Conference (TREC-3), pp. 69-80, 1995.
3) I. Dagan, S. Marcus and S. Markovitch: Contextual Word Similarity and Estimation from Sparse Data, Proc. of 31th Annual Meeting of the Association for Computational Linguistics, pp.164-171, 1993.
4) E. Keen: The Use of Term Position Devices in Ranked Output Experiments, The Journal of Document, Vol. 47, No.1, pp.1-22, 1991.
5) D. Harman editor: The 1st Text Retrieval Conference (TREC-1), National Institute of Standards and Technology, 1993.
6) D. Harman editor: The 2nd Text Retrieval Conference (TREC-2), National Institute of Standards and Technology, 1994.
7) D. Harman editor: The 3rd Text Retrieval Conference (TREC-3), National Institute of Standards and Technology, 1995.
8) G. Salton and M. McGill: Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.
9) G. Salton: Automatic Text Processing - The Transformation, Analysis and Retrieval of Information by Computer, Addison-Wesley Publishing, 1989.
10) A. Singhal, G. Salton, M. Mitra and C. Buck-

ley: Document Length Normalization, Technical Report 95-1529, Cornell University, 1995.
11) A. Singhal, C. Buckley and M. Mitra: Pivoted Document Length Normalization, Proc. of 20th ACM SIGIR, pp. 21-29, 1996.
12) A. Singhal: AT&T at TREC-6, In E. Voorhees and D. Harman, editors, Proc. of The 6th Text Retrieval Conference (TREC-6), pp. 215-226, 1998.
13) E. Voorhees and D. Harman: Overview of the Sixth Text REtrieval Conference (TREC-6), In E. Voorhees and D. Harman, editors, Proc. of The 6th Text Retrieval Conference (TREC-6), pp. 1-24, 1998.
14) E. Voorhees and D. Harman: Overview of the Seventh Text REtrieval Conference (TREC-7), In E. Voorhees and D. Harman, editors, Proc. of The 7th Text Retrieval Conference (TREC-7), pp. 1-23, 1999.
15) R. Wilkinson: Effective Retrieval of Structured Documents, Proc. of 17th ACM SIGIR, pp. 311-317, 1994.
16) D. Yarowsky: Word-sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora, Proc. of COLING-92, pp.454-460, 1992.
17) J. Xu and W. Croft: Query Expansion Using Local and Global Document Analysis, Proc. of 20th ACM SIGIR, pp. 4-11, 1996.
18) 木谷 強ほか: 日本語情報検索システム評価用テストコレクション BMIR-J2, 情報処理学会研究会報告, Vol. DBS114-3, pp.15-22, 1998.
19) 木谷 強: 固有名詞の特定機能を有する形態素解析処理, 情報処理学会研究会報告, Vol. NL90, pp.73-80, 1992.
20) 木本晴夫ほか: 情報検索システム評価用データベースの構築の提案, 情報処理学会研究会報告, Vol. FI32-1, pp.1-8, 1993.
21) 神門典子: 情報検索システムの評価を巡って テストコレクションとコンペティションを中心に, 1999 年情報学シンポジウム講演論文集, pp.129-136, 1999.
22) 芥子育雄ほか: 情報検索システム評価用ベンチマーク Ver1.0 (BMIR-J1) について, 情報処理学会研究会報告, Vol. DBS106-19, pp.139-145, 1996.
23) 関根 聡, 井佐原均: IREX:情報検索, 情報抽出コンテスト, 情報処理学会研究会報告, Vol. FI-51-15, pp.109-116, 1998.
24) 高木 徹, 木谷 強: 単語出現共起関係を用いた文書重要度付与の検討, 情報処理学会研究会報告, Vol. FI-41-8, pp.61-68, 1996.
25) 松本裕治, 北内啓, 山下達雄, 平野善隆, 今一修, 今村友明:「茶筌」version1.0 使用説明書, NAIST Technical Report, NAIST-IS-TR97007, 1997.
26) 武藤眞介: 統計解析ハンドブック, 朝倉書店, 1995.

(平成 11 年 6 月 20 日受付)
(平成 11 年 9 月 27 日採録)

高木 徹（正会員）
1990 年筑波大学第三学群情報学類卒業. 1992 年同大大学院修士課程理工学研究科修了. 同年ＮＴＴデータ通信 (株) (現 (株) ＮＴＴデータ) 入社. 情報検索等の研究開発に従事.

木谷 強（正会員）
1960 年生. 1983 年慶應義塾大学工学部電気工学科卒業. 同年日本電信電話公社 (現ＮＴＴ) 入社. 1988 年 NTT データ通信 (株) に転籍. 形態素解析, 情報抽出, 情報検索に関する研究開発に従事. 現在、(株) ＮＴＴデータ技術開発本部北米技術センタ部長. 博士 (工学). 平成 10 年度坂井記念特別賞受賞.