自然言語処理　144－10
（２００１．７．１６）

学習分野別の共起語情報を用いた学習情報の検索手法の検討

鈴木 雅実† 松本 一則‡ 井ノ上 直己†‡ 橋本 和夫‡ 中山 実* 清水 康敬**

†通信・放送機構 ‡KDDI研究所

*東京工業大学 **国立教育政策研究所

E-mail : admin@coop-m.central.ed.tao.go.jp, msuzuki@kddlabs.co.jp

あらまし　インターネット上の多様な情報源へのアクセスが可能となった今，教育・学習に利用できる情報（コンテンツ）を容易に検索することを支援する機能が求められている。本研究では，学習分野野ごとに整理された学習関連のWebページに含まれる語の共起情報を用いて，利用者の検索語に新たな検索語を追加する手法を提案する。理科教育の下位分野（天文・物理・化学等のサブカテゴリー）毎に抽出した共起情報を用いて検索語を２～３語追加することにより，新聞記事を対象とした検索実験において精度を高めることができた。この傾向は，検索結果の上位において顕著であり，効率良く学習情報を検索するための支援方法として有望である。

キーワード　学習情報，文書検索，共起語，検索質問の拡張，類似(連想)検索

Document Retrieval based on Word Co-occurrences extracted from Educational Information Resources on the Internet

Masami Suzuki†, Kazunori Matsumoto‡, Naomi Inoue†‡, Kazuo Hashimoto‡,
Minoru Nakayama* and Yasutaka Shimizu**

†Telecommunications Advancement Organization of Japan
‡KDDI R&D Laboratories Inc.
*Tokyo Institute of Technology
**National Institute for Educational Policy Research

E-mail : admin@coop-m.central.ed.tao.go.jp, msuzuki@kddlabs.co.jp

Abstract　Currently it is difficult to easily find usable educational contents from a lot of various information resources on the Internet. In this article, we propose a new method to expand the user's query, based on word co-occurrences extracted from classified educational documents. In our recent experiment using newspaper articles, a single-word query with several additional major co-occurrent words showed better result with higher precision, in a case of 7 subcategories for high school science domain. This tendency seems prominent in lower recall part of the results, and will contribute to efficient educational information retrieval.

key words　Educational Information, Document Retrieval, Word Co-occurrence, Query Expansion, Associative Document Search

1. はじめに

インターネットの教育利用の活発化に伴い，膨大な情報の中から適切な情報を取捨選択することは，生徒だけではなく教師にとっても重要な技法となりつつある [1]。しかし，利用者個人の処理能力には限界があるので，効率的な学習情報検索のための技術的な支援方法がより重要になると考えられる。すなわち，曖昧になりがちな検索要求（検索意図）の明確化をを助ける手段が必要である。さらに，検索結果の一覧から有用と思われる情報（文書）を迅速に発見する手段が必要である。これらには，検索精度の向上が求められる[1)]。

検索要求の明確化のために，元の検索要求に情報を付加するなどの方法は，検索要求の拡張（Query Expansion）として研究されている。この手法は，以下の２通りに大きく分類できる。

(1) 検索結果のフィードバック
検索結果に対する利用者の適合度判断を反映させる方法や，上位の結果を無条件で用いる擬似フィードバック手法が知られている。ただし，利用者の負担や，判断に寄与する情報の信頼性，検索式の重みづけ等の問題が指摘されている [2]。

(2) 関連語情報の提示
利用者に対して参考となる関連語情報を，外部の文書群やシソーラス・辞書等の知識ベースから抽出して提示する手法が提案されている。例えば，大量文書から抽出した共起情報を検索要求の拡張手段の一つとして適用する情報検索システムの試作事例（[3]など）がある。大量の情報提示が必要となるため，ユーザ・インタフェースに主眼が置かれる傾向がある。

本研究では，利用者からのフィードバック情報は重要と考えるが，それを経験的な知識として再利用するために，上記の（2）のアプローチを採ることにした。これまで，教育や学習情報に関しては，この種の知識として蓄積された事例はほとんどない。小学校の学習指導要領を用いて検索語の追加を支援するシステムが見られる程度である [4]。また，本研究が対象とするようなインターネット上の学習情報については，ほとんど検討されていない。

そこで，本研究では，実際にインターネット上で，学習利用可能と判断された文書集合を分析し，検索要求の拡張に利用可能な知識としての共起語情報を抽出する。そして，これを与えられた検索語と関連する語の追加方法に適用して，検索効率への効果を検討することにした。

なお，本研究で語の共起情報に注目する理由は，ある用語から連想的に想起される共起語関係が，その用語を中心とする特定の学習内容に深く関わると考えられるからである。また，ベクトル空間モデル等において，類似する単語ベクトルを持つ類義語を検索語として加えた場合に，再現率は高くなるが必ずしも検索精度の向上に寄与しない点も考慮した。

このように，本研究では，検索精度を重視するが，検索語を追加した上で次に述べる類似文書検索を実行することにより，再現率を保つようにしている。

与えられた文書と内容的に類似した文書を，検索対象の中から発見する類似文書検索（または連想検索 [5]）手法は，検索要求を明示的に記述することが困難な場合に役立つ技術である。この手法では，検索要求となる文書および検索対象となる文書群に含まれる語の分布の相関に着目し，類似度の高い文書を順序づけて提示することが可能である。種々の文書間類似度計算モデルがあるが，本研究では文書クラスタリングにおいて精度が良いとされる確率型モデル [5]を採用している[2)]。

一方，類似文書検索においては，検索要求となる断片文や一まとまりの文書から抽出される検索語の集合が，検索意図を十分に反映していることが求められる。しかし，検索要求に含まれる情報自体が不足しているような場合は効率的とは言えない。そこで，利用者の学習情報を検索する意図を補完するために用いるのが，前述した共起語情報である。

本研究の目的を以下に示す。

(1) 学習可能な文書内における語の共起関係を分析する。この結果を適用した共起語情報に基づく検索語の追加手法を開発する。

(2) 検索語の追加による検索精度の向上を確認するために，新聞記事をテストデータとして，本手法による検索性能の評価を行なう。評価結果より本手法の効果を定量的に明らかにする。

(3) 提案する手法を組み込んだ検索システムの実用上の問題点等を検討する。

これらの目的のために，中学校から高等学校レベルの理科についてのWeb 学習情報を選択し，これらの学習情報源での語の共起情報を抽出して，具体的な検索語の追加について検討した。

2. 共起語情報を用いた検索語の追加手法

2.1 検索語の追加方法の概要

ここでは，まず具体例として，後述する評価実験

1) このほか，検索結果の一覧において文書の選択を容易とする情報提示方法の改善も重要であるが，本稿での議論の対象外とする。

2) 同モデルを用いた特許文書の類似検索実験では，国際特許コードや通常のキーワード入力によるブーリアン型の検索と比較して，専門家の目で見原文書と類似性の高い文書が上位の（類似度の高い）検索結果として得られ易い傾向が確認されている [6]。

で対象とした「理科」の教科領域内での検索を想定して，概要を説明する。例えば，天文関連の用語である「火星」に関する学習情報を検索する場合を仮定する。「火星」のみを検索語とした場合は，検索意図が曖昧で，検索範囲によっては大量の不適合文書が検索されてしまう。もし，これに学習利用可能と判断された文書群内で「火星」と高頻度で共起する語を追加したとすると,より検索結果は絞り込まれたものとなる。しかし，通常のキーワードのAND検索では絞り込みによる検索漏れが生じ，OR検索では発散した検索結果となり易い。そこで，検索語を追加した上で，類似文書検索を実行すれば，精度と再現率の両立を図ることが可能と考えられる。

以上のような考察に基づき，学習分野毎に抽出した共起語情報を追加する検索手法のアウトラインを示したものが，図１である。まず，学習情報源における語の共起情報を収集する。図１の中央に示すように，あらかじめ定めた分野ごとに語の共起情報を抽出する。この情報から分野ごとに出現する任意の語に対する共起語リストが，後述する共起度順に取得できるようにする。これを検索拡張に利用可能な知識とする。このように構成された知識を用いた検索手順を以下に示す。

(1) 検索要求の入力
検索要求としての検索語1語と，検索対象とする分野を指定する。

(2) 学習分野の指定と共起情報の参照
入力された検索語と指定された分野における共起語テーブルを参照し，検索語の追加を行なう。追加語数については，後述する。

(3) 検索の実行
追加された検索語を含む，拡張検索要求に基づいた類似文書検索処理を実行する。

なお，図１の中で点線で示した部分は，将来的に，提示された検索結果の中から利用者が有用と判断した文書を，共起情報を抽出するための活用事例データにフィードバックすることを想定している。このように，循環的な性能向上を図る機構を視野に入れたものである。また，利用者グループの間で，実際に学習利用向けと判断された文書を，分野別の事例として蓄積することにより，検索に関わる知識を共有するような枠組みとすることも可能である。

2.2 学習情報源からの共起語情報の抽出

本研究では，インターネットでよく検索される中学・高校で学習する理科の内容について検討を行なった。中学校や高校で指導される理科の内容は，表１の左欄に示すような７分野からなる。これらの分野で利用可能なWeb学習素材を「理科」に関係する教育・学習リンク集等を基に合計約4,200ページを収集した。そして，収集したWeb学習素材を，人手によって先の７分野に分類した。各分野におけるページ数は表１の通りである。これらのWebページのテキスト部分の名詞に着目し，各分野毎の語の共起頻度を抽出した。抽出のための語の切り出しは，形態素解析ツールChasen 1.5を用いた。また，共起頻度はページ内での同時出現度数として集計した[7]。

2.3 共起語情報を用いた検索語の追加

前項で述べたように，指定された学習分野での共起語情報の参照により，検索語を追加する手順について述べる。

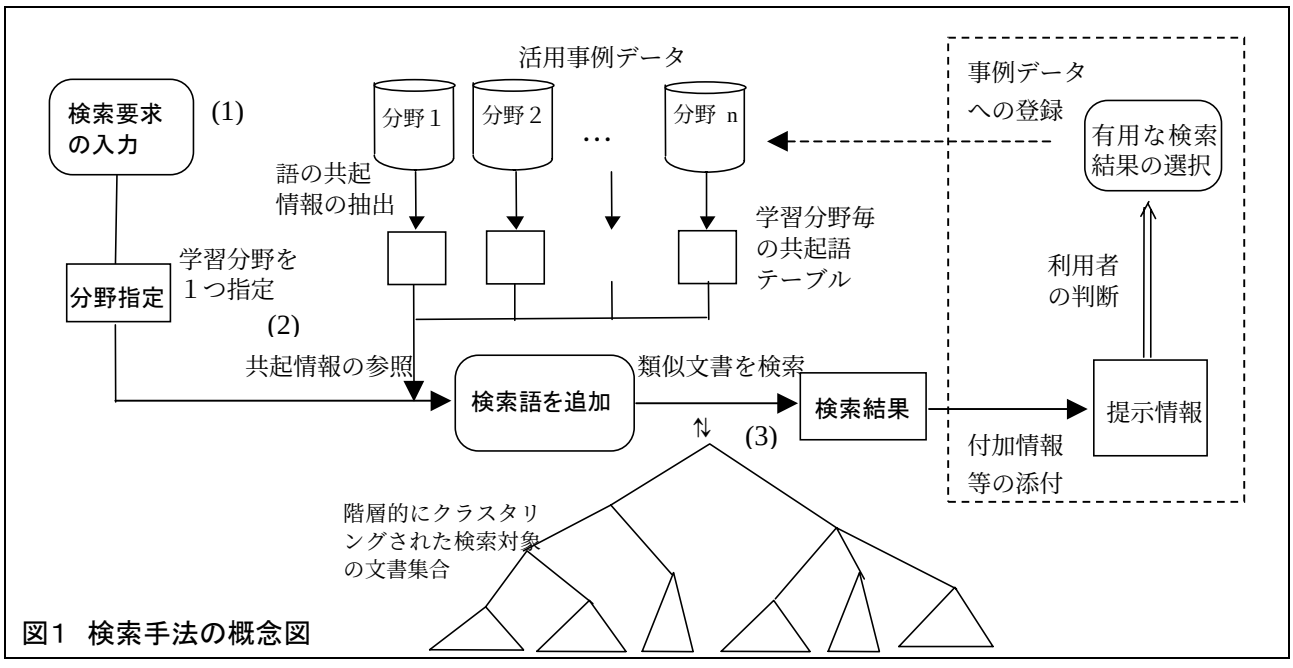

図1 検索手法の概念図

表１　学習分野と収集したページ数

学習分野	ページ数	代表語（頻度順）
天文（UNI）	722	観測，星，月，日，天文
地学（EAR）	342	火山，岩，石，地球，堆積
物理（PHY）	338	実験，光，物理，法則，製作
化学（CHE）	456	反応，水，実験，結合，分
動物（ANI）	680	メダカ，章，宇宙，実験
植物（PLA）	1,009	図鑑，花，皮，樹木，植物
気象（WEA）	667	気象，雨，気温，用語
合 計	4,214	

単純に共起頻度の高い語を順に選ぶ方法では，元の語と共起する語双方の単独での出現頻度に対する，相対的な共起性が反映されない。そこで，ここでは次の表２に示すようにある語と別の語との共起が偶然で独立性が高いか，偶然でなく依存性が高いかの比率の差分によって，共起度を定義する。

共起を収集する文書集合を D $\{d_1, d_2, \cdots, d_N\}$，集合内の異なり語を$\{w_1, w_2, \cdots, w_M\}$ とする。ここで，d_iは個々の文書を，w_jは個々の異なり語を指す。

共起度数 C は，

$C(d_i, w_j, w_k) = 1$ if w_j と w_k が同一文書に出現している

$= 0$ if w_j と w_k が同一文書に出現していない

ここで，前述した相対的な共起性による頻度分布を考慮した共起度数 Cooc を求める。

表２のような２×２分割表と，その観測値が与えられた際の尤度計算方法については，文献[8]で述べられている。これを参考に，２語の共起を，独立モデルと依存モデルの対数尤度の差分と考えた場合の共起度 Coocを次のように定義する。

$$\begin{aligned}\mathrm{Cooc}(w_j, w_k) &= LL(n_{11}, n_{12}, n_{21}, n_{22}) \\ &= (n_{11}+n_{12})\log(n_{11}+n_{12}) \\ &+ (n_{11}+n_{21})\log(n_{11}+n_{21}) \\ &+ (n_{21}+n_{22})\log(n_{21}+n_{22}) \\ &+ (n_{12}+n_{22})\log(n_{12}+n_{22}) \\ &- N\log N \\ &- (n_{11}\log n_{11} + n_{12}\log n_{12} + \\ &\quad n_{21}\log n_{21} + n_{22}\log n_{22})\end{aligned}$$

ただし，$N = n_{11}+n_{12}+n_{21}+n_{22}$

ここで，マイナスの共起を取り除くため，次の条件をつける。

$$\frac{n_{11}}{n_{11}+n_{21}} > \frac{n_{12}}{n_{12}+n_{22}}$$

表２　共起語同士の出現頻度

	w_k	$\neg w_k$
w_j	w_j，w_kが出現する文書の数 n_{11}	w_j が出現し，w_kが出現しない文書の数 n_{12}
$\neg w_i$	w_j が出現せず，w_k が出現する文書の数 n_{21}	w_j，w_k ともに出現しない文書の数 n_{22}

この条件を満たすものだけを，（プラスの）共起語と認定する。

以上のように定義した対数尤度に従って，ある語に対する共起語は順序づけられる。これを共起度と呼ぶことにする。この共起度の順序に追加する検索語を選ぶことができる。表３に，各分野の代表的な語について，共起度順に追加すべき検索語の候補例を示す。

表３　共起度に基づく検索語の追加例

学習分野	入力検索語	追加する検索語例
UNI	光年	＋距離，銀河系，恒星，銀河
EAR	大陸	＋岩，地殻，地球，体
PHY	エネルギー	＋光，実験，運動，法則
CHE	炭素	＋結合，電子，反応，分子
ANI	卵	＋細胞，生物，発生，実験
PLA	樹木	＋図鑑，皮，発生，花
WEA	気温	＋西，平年，気象，予報

3. 評価実験

3.1 評価方法

提案手法を評価するため，次のような検索実験を行なった。本実験では検索語1語の性能に対して，検索語の追加語数による検索精度・再現率等の性能を調べた。本評価で対象としたのは新聞記事１年分である。今回の実験で新聞記事を用いた理由は次の通りである。

- 検索実験に必要な比較的多数の文書集合として利用可能であり，均質性が高い。
- 新聞記事に教育向け素材を含んだものも多く，教育目的で利用される場合も多い。。
- 検索語は含んでいても教育利用には適さないもの，語の多義性により検索目的と合致しないもの等を包含する文書集合であり，検索精度等の評価を実行する対象としてふさわしい。

評価実験では，研究用に供与されているA新聞電子化記事のうち1998年分より約2万件を使用した。これらの記事から学習利用可能な情報を含む記事を人手で判断し，正解集合として用意した。実験に用いた検索要求としての検索語は，各学習分野毎

に出現頻度が比較的高い語群から任意に選んだ。すなわち，表４に示す７分野の各３語である。

検索実験においては，前述の方法で検索語の追加を行なった。追加する共起語を選択する情報源となるWeb文書は，表１に示す，総計約4,200ページである。たとえば，検索語「光年」に対しては，「距離」，「銀河系」，「恒星」，… の順に追加した。また，比較のために学習分野を限定せずに全分野から共起語を選択して検索語を追加した場合も同様に評価した。たとえば，検索語「火星」に対しては，分野「天文(UNI)の共起情報を指定した場合と，分野共通の全体共起情報を使用した場合との２種類の実験結果が導かれる。検索結果については，累積の再現率に対する平均精度を計算する，TREC_EVAL [9]を利用して，傾向を分析した。

3.2 実験結果

21個の検索語について行なった実験結果を集計し，再現率と検索精度の関係としてグラフ化した結果を図２と図３に示す。図２は，学習分野を限定して共起語を選択した場合である。また．図３は学習分野を限定せずに共起語を選択した場合の結果である。これらのグラフで「検索語のみ」は，与えられた検索語のみで検索したものである。「共起語＋１」「共起語＋２」… は，それぞれ検索語

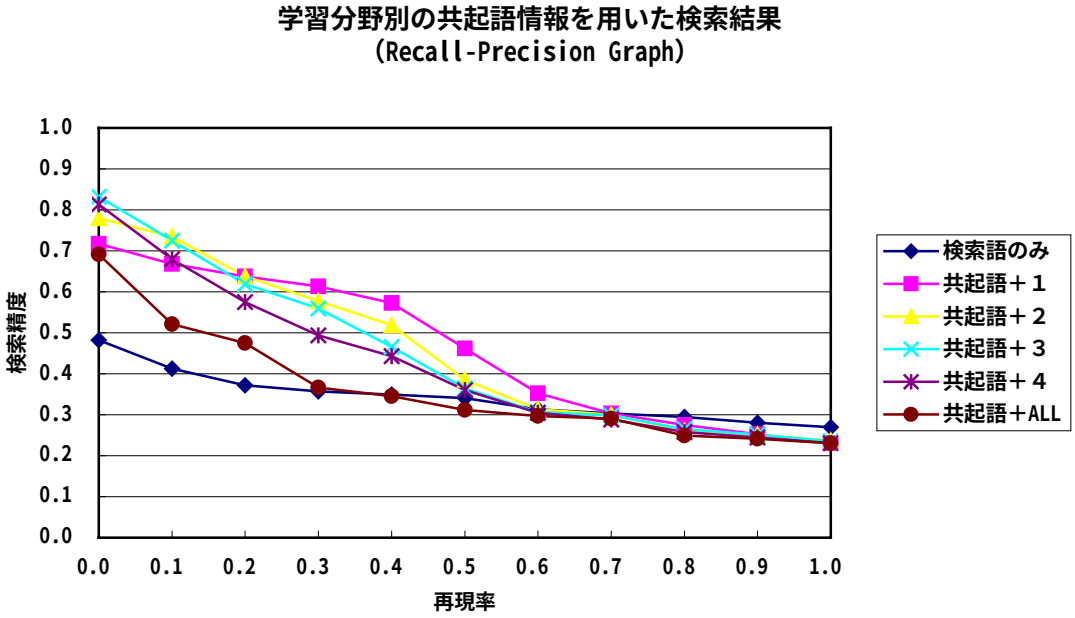

図２ 学習分野を限定して共起語情報を用いた場合の検索結果

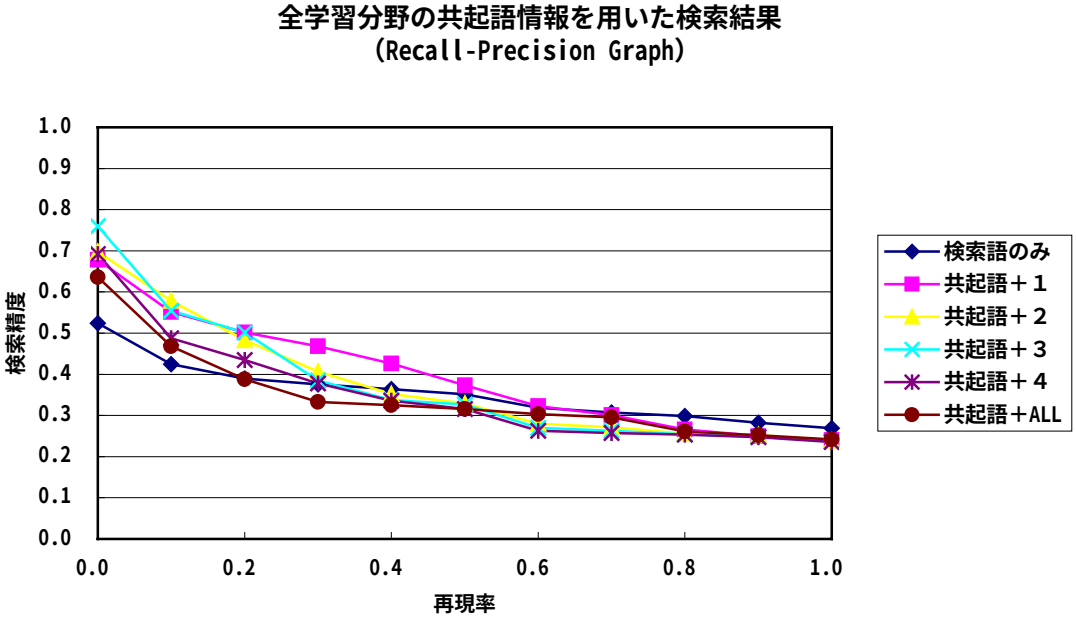

図３ 学習分野（理科）全体の共起語情報を用いた場合の検索結果

表４　実験に用いた検索語の一覧
（括弧内は正解記事数）

学習分野	検索語
UNI	彗星(9)，火星(44)，光年(51)
EAR	化石(162)，プレート(35)，カルデラ(14)
PHY	エネルギー(429)，振動(38)，レーザー(41)
CHE	炭素(23)，硫酸(5)，有機化学(6)
ANI	産卵(168)，交尾(16)，コウモリ(27)
PLA	子葉(4)，胞子(12)，広葉樹(53)
WEA	高気圧(44)，梅雨(244)，雷(32)

と共起度の高い語を１個，２個，…と追加した場合の結果である。また，「共起語＋ALL」は，共起度の尤度値が一定値（信頼性の観点から10に設定）以上の語をすべて追加して検索した結果を示す。図２では，再現率が低い条件，すなわち検索結果の上位として示される部分において，共起語を追加した場合の検索精度の向上が顕著であることが分かる。また図３でも同様の傾向が見られる。しかし，学習分野を限定して共起語を選択した場合の図２ほど顕著な精度向上は見られない。以上の実験値に対して，追加共起語数の違いをパラメタとする，各系列間での分散分析による検定を行ない，次のような観察結果が得られた。

(1) 共起語の追加に対する検索精度の変化
　学習分野限定および全学習分野のどちらの共起情報を参照した場合も，共起語を２～３語追加した際に最も検索精度が向上し，さらに追加して行くと精度は低下する。
・分野別に限定した共起語情報を参照した場合では，「検索語のみ」と「共起語＋１」「共起語＋２」「共起語＋３」「共起語＋４」の間で各々有意差が確認された（$p<.01$）。また，「共起語＋１」～「共起語＋３」は「検索語＋ALL」と比較して有意に精度が高かった（$p<.01$）が，それらの間では有意差は確認できなかった。
・分野全体の共起語情報では，「共起語＋１」が「検索語のみ」「共起語＋４」「共起語＋ALL」と比較して有意に精度が高かった（$p<.01$）。

(2) 参照する共起情報の取得範囲による比較
　学習分野別の共起情報を用いた方が，全学習分野から抽出した共起情報を参照する場合よりも，高い検索精度が得られる。
　・両者の「共起語＋２」および「共起語＋３」の系列間で，有意差が見られた（$p<.01$）。
　・この傾向は，検索対象文書数が多い場合に一層顕著であった。

(3) 対象文書数の大小による結果の比較
　上記の(2)に関して，単純に検索語が含まれる対象文書数の大小（400記事以上／以下）による２分比較を行なった。この結果，対象文書数が大きい検索語については，学習分野別の共起情報を用いた場合は，精度の向上効果が顕著である。これに対し，全分野の共起情報を用いた場合は，精度はそれほどは向上しない。
　一方，これと反対に，対象文書数が小さい検索語については，精度の向上は比較的緩やかである。これを図４に示す。すなわち，検索語のみの条件から，共起語を１語追加した場合の精度の向上率は，対象文書数が大で分野別の共起語情報を用いた場合は，２倍程度に達している（図４上の右側のグラフを参照）。なお，共起語の追加による精度の向上は，再現率の低い部分（検索結果の上位）で顕著なことから，再現率＝0.5までの平均精度で比較を行なった。

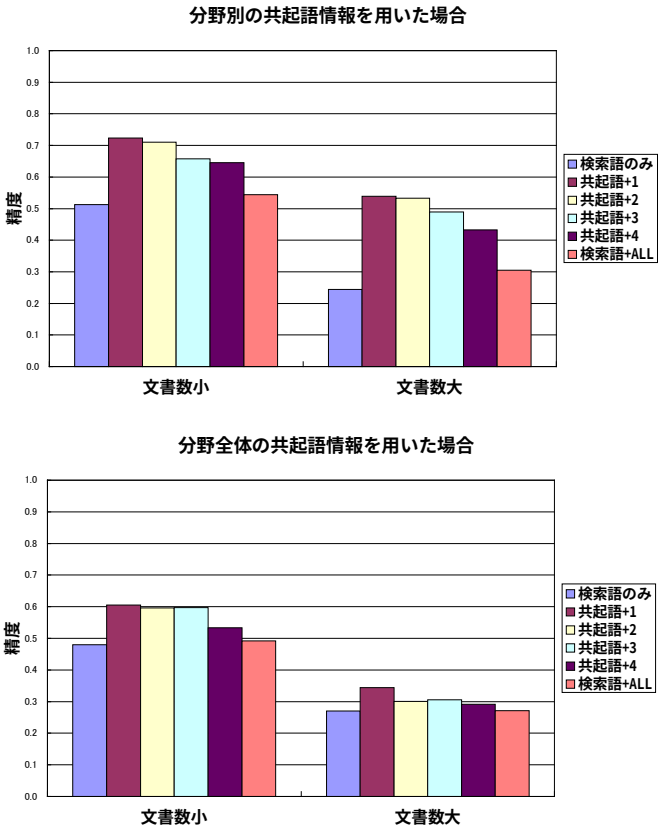

図４　対象文書数の大小による比較

注）対象文書数の大小は400文書以上／以下
　上図は分野限定の共起語情報を使用
　下図は分野共通の全体共起語情報を使用
　再現率＝0.5までの平均精度で比較

(4) 典型文書の検索結果
　新聞記事と比較するために，各検索語について典型的な学習向け文書を作成して，検索を行なった。典型文書の作成には，学習参考書・事典等の資料を用いた。21の検索語について各3文書づつを用意し，新聞記事集合に追加したものを検索対象の文書集合として実験した。この結果，計63文書について，分野別の共起語情報および分野共通の全体共起語情報を参照した場合のいずれも，共

起語＋２の場合が平均して最も高い検索順位を示した。新聞記事の場合と対照するため，検索結果の上位20位以内の再現率として集計したものが，次の表５である。すなわち，正解文書のうちで上位20位以内の検索順位となった文書の割合を示している。上位20位以内というのは，サーチエンジン等で，検索結果の一覧の最初のページとして表示される件数にほぼ相当する。表５を見ると，サンプル数が少ないことを考慮に入れても，典型的な学習向け文書の場合は，検索結果の上位に比較的よく現れ易いことが分かる。このことも，本研究の手法の利用可能性を示唆している。

表５　検索結果上位 20 位内の再現率

	新聞記事集合	典型文書
検索語のみ	0.08 0.09	0.18 0.15
共起語＋１	0.13 0.11	0.55 0.40
共起語＋２	0.14 0.10	0.63 0.48
共起語＋４	0.12 0.08	0.60 0.45
共起語＋ALL	0.08 0.07	0.47 0.47

注) 上段は分野別の共起語情報を使用した場合，下段は分野共通の全体共起語情報を使用した場合の平均の再現率を表す.

(5) その他の観察事項

類似文書検索の結果は，結果検索対象の文書のテキスト長によって影響される。本実験で検索対象としている新聞記事を，そのサイズにより長文および短文の記事集合として２分して実験を行なった場合の結果は，短文記事集合の方が長文記事集合よりも，平均して１割程度高い検索精度となった。この問題に関しては，Singhalらによる補正方法が提案されている [10]。

また，今回の評価実験では，共起語情報を抽出する範囲を同一文書内としたことにより，内容とは関係なく出現する，ノイズ語が共起度の高い語となる場合が一部見られた。従って，共起語情報を抽出する範囲を，ページ内よりも局所的に限定した，ある程度近接した範囲内から抽出する方法が有効と考えられる。

4. 考察

4.1 実験結果に関する検討

２～３語共起語を追加することにより，検索精度が向上することは，直観と矛盾せず，期待した効果を示すものである。さらに共起語を追加して行くと平均の検索精度が低下するのは，次のような理由によるものと考えられる。すなわち，共起語テーブルで与えられる情報は，元の検索語に対する共起度の順位であり，それらの共起語群が実際に同一文書内に現れる確率は，語数が多くなるほど少なくなると考えられるからである。そこで，実際の正解集合全体に対して最も類似度が高くなる極大点が，２～３語共起語を追加した場合ということになる。さらに，再現率の中間値0.5の前後で比較した場合，前半（検索結果上位）における検索精度の向上が，後半（検索結果下位）よりも顕著であることは，本手法が，１章で述べたように検索精度を重視する検索において優れた効果を示すものである。

また，この実験で対象とした「理科」の学習領域内でも，下位分野毎に抽出した共起語情報が，分野全体の共起語情報よりも，多くの場合に高い検索精度を導いたことから，固有の学習分野毎に共起語情報を抽出して利用することの有効性が示されたと言える。さらに，本手法による共起語追加が検索精度の向上に特に寄与するのは，検索対象の文書数が大きい場合であることが明らかとなった。このことは，キーワードのAND条件検索による，絞り込み検索で精度が向上する場合と相似した効果と考えられる。

4.2 他の手段との比較

類似文書検索を用いずに，単純なAND検索やOR検索を行なった検索結果を表６に示す。この比較実験では，追加する共起語を今回の実験と同等として，汎用の全文検索システムNamazu [11] を用いた。検索実験実験対象の文書集合は，３章で用いた新聞記事である。また，AND検索・OR検索については，各検索語について得られる再現率および精度の値の組のマクロ平均をとったものである。表内のRは再現率を示す。なお，本手法の値は再現率R＝0.5までの平均精度を比較として記した。

表６　AND ・ OR 検索との比較

	本手法	AND 検索	OR 検索
検索語のみ	0.39 0.40	0.28 0.28	0.28 0.28
共起語＋１	0.61 0.50	0.57 (R=0.19) 0.49 (R=0.24)	0.37 (R=0.16) 0.32 (R=0.17)
共起語＋２	0.61 0.47	0.66 (R=0.08) 0.63 (R=0.13)	0.39 (R=0.25) 0.34 (R=0.31)
共起語＋３	0.59 0.48	—	0.32 (R=0.27) 0.35 (R=0.39)
共起語＋４	0.56 0.44	—	0.38 (R=0.31) 0.35 (R=0.45)

注) 上段は分野別の共起語情報を使用した場合，下段は分野共通の全体共起語情報を使用した場合の平均の精度を表す. 本手法の値は再現率R＝0.5までの平均精度.

これを見ると，AND検索では，共起語を追加した場合に，絞り込み効果により検索精度が向上するが，同時に再現率が低下することが分かる。一方，OR検索の場合は，再現率は保たれる傾向にあるが，共起語を追加した場合も検索精度はあまり変化しない。従って，総合的に見た場合，共起語情報を用いる場合に，類似文書検索を適用した方が，精度と再現率を両立させる意味で効果がある。

なお，これに加えて，元の検索語に追加する語を他の方法で選んだ場合との比較実験を予定している。

5. まとめ

学習情報を効率良く検索するため，学習分野の共起語情報を用いた検索手法について，その有効性を確認するための実験結果を中心に報告した。実験では，中学校～高校レベルの理科の学習領域内で，７つの下位分野毎の学習利用可能な文書群から抽出した共起情報を用いて，元の検索語に検索語の追加を行なった。以下に結論をまとめる。

(1) 分野別に収集した学習利用可能な文書集合から，出現分布を考慮した共起語情報を抽出した。これを基に，検索語に対して，指定された分野内の共起度の強い語を順に追加して，類似文書検索を実行する機能を実装した。

(2) 新聞記事を対象とした検索実験の結果，共起語を２～３語追加した場合に，最も検索精度が向上することが確認できた。また，学習分野（７つの下位分野）別の共起情報を用いる方が，全学習分野（理科全体）から得られた共起情報を用いる場合と比較して，より精度が向上することが分かった。さらに，AND・OR検索との比較では，検索精度と再現率の双方を追求する上で，利点があることが確認された。

(3) 典型文書を用いた検索実験結果等を合わせて考察すると，本研究で提案した共起語情報の追加を行なう検索手法は，学習利用可能な文書を効率良く検索するための支援方法として，貢献できるものと考えられる。

また，本手法の利用面を考えると，検索精度を向上させる可能性の高い検索語を含む追加語候補を提示して，利用者に選択してもらうような，検索インタフェースを用いたインタラクティブな検索を指向することも可能である。

さらに，学習情報の検索に適した質の高い共起語抽出方法や，体系的な知識源との統合利用等に関しては，今後の検討課題である。

謝辞

本研究は，通信・放送機構(TAO)の直轄研究「学校における複合アクセス網活用型インターネットに関する研究開発」の一環として実施しているものである。関係各位の支援と助言に感謝する。

参考文献

[1] 越桐國雄：学校教育におけるインターネットの利用 ―現状と展望― ，情報処理，Vol.42 No.1, pp. 58 – 63, 2001.

[2] 徳永健伸：“情報検索と言語処理”，辻井潤一編，言語と計算－５，東京大学出版会，1999.

[3] 高山泰博・R. Flournoy・S. Kaufman・S. Peters : “単語間の連想関係に基づく情報検索システムInfoMAP”，情報処理学会研究会報告，FI53-1, 1999.

[4] 森本容介・中山実・清水康敬：“学習用Web情報の検索支援システム”，教育システム情報学会誌，Vol.17 No.3, 2000.

[5] 岩山真・徳永健伸：“確率的クラスタリングを用いた文書連想検索”，自然言語処理，Vol.5 No.1, pp.101 –118, 1998.

[6] N. Inoue，K. Matsumoto，K. Hoashi and K. Hashimoto : “Patent Retrieval System Using Document Filtering Techniques”, *ACM-SIGIR 2000 Workshop on Patent Retrieval*, 2000.

[7] 鈴木雅実・松本一則・井ノ上直己・橋本和夫・中山実・清水康敬：“学習情報源に含まれる語の共起を用いた文書検索手法の検討”，情報処理学会第62回全国大会，2000.

[8] 鈴木義一郎:“情報量基準による統計解析入門”，講談社サイエンティフィク，1995.

[9] Text REtrieval Conference (TREC)：http://trec.nist/gov/

[10] Singhal, A., Buckley, C. and Mitra, M. : “Pivoted Document Length Normalization”, *SIGIR96*, 1996.

[11] 全文検索システムNamazu：http://www.namazu.org/