1X-5 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

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† KDD研究所‡ 東京工業大学*

1. はじめに

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ている現在，教育・学習に利用できる情報（コンテンツ）を容易に検索するシステムが求められている。本研究では，学習分野ごとに整理された教育・学習関連の Web ページに含まれる語の共起情報を基に，検索者の検索語に新たな検索語を追加する手法を提案する。本稿では，理科教育の分野での共起情報を用いて検索語の追加例を報告する。

2. 背景と目的

2.1 類似文書検索手法の利用

類似した文書を，検索対象の中から発見する類似（または連想）文書検索手法は，通常のサーチエンジンなどのブーリアン型キーワード検索よりも優位であることが知られている。この手法では，該当文書に含まれる語の分布（出現頻度，共起性）に着目する。種々の文書間類似度計算モデルがあるが，本研究では確率型モデルを採用した [1]。同モデルを用いた新聞記事の文書クラスタリング実験で，政治／経済／社会／外報／運動／学芸の 6 分野で各 100 件の記事を用いた分類性能を報告した [2]。この手法は，本稿で述べる新たな提案においても利用する。

効率的な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，一般的な検索語を入力する文書検索においては，入力した検索語だけでは情報が不十分である場合が多い。この場合，類似検索手法を用いても検索精度はあまり期待できない。

2.2 共起情報に基づく検索語の追加方法

たとえば「人類の誕生」といった断片文の中から「人類」「誕生」を検索語として検索しても，検索結果の文書集合の中から学習に適した素材を探すことは容易ではない。そこで，学習情報源に含まれる語の共起情報を利用して，検索語を追加する拡張（あるいは補完）を行なった上で類似（連想）検索を行えば，より精度の高い検索結果が得られると期待される。

このような観点から，共起情報を考慮した情報検索システムの試作事例がある[3]。また，教育・学習の情報源として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手法も提案されている[4]。

ところで，学習に利用可能な文書素材と言っても，学習分野毎に含まれる語や語の分布傾向は異なる。また，含まれる語の共起頻度も各々の特徴を反映していると考えられる。従って，入力される検索語に該当する分野毎に，それぞれ蓄積された語の共起情報を参照した検索語を追加する方法が考えられる。

以上のような検索手法のアウトラインを次ページの図1に示す。この図のように，検索要求の入力とともに学習分野を指定すると，その分野に応じた共起情報が参照され，検索語の追加が行なわれた後に，類似検索を実行する。将来的には，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活用事例データにフィードバックするような，循環的な性能向上も視野に入れている。

3. 共起データの収集

提案手法に必要な学習分野における語の共起情報を得るために，以下のように文書を収集分析した。中学校ー高校レベルの「理科」において7つのサブカテゴリーを設定し，インターネット上の学習 Web 素材を収集した。実際に取得したページ総数は，表1のように合計約 4,200 ページである。これらの収集ペ

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

Document Retrieval based on Word Co-occurrences extracted from Educational Information Resources on the Internet
Masami Suzuki (Telecommunications Advancement Organization of Japan), Kazunori Matsumoto, Naomi Inoue, Kazuo Hashimoto (KDD R&D Laboratories, Inc.), Minoru Nakayama and Yasutaka Shimizu (Tokyo Institute of Technology)

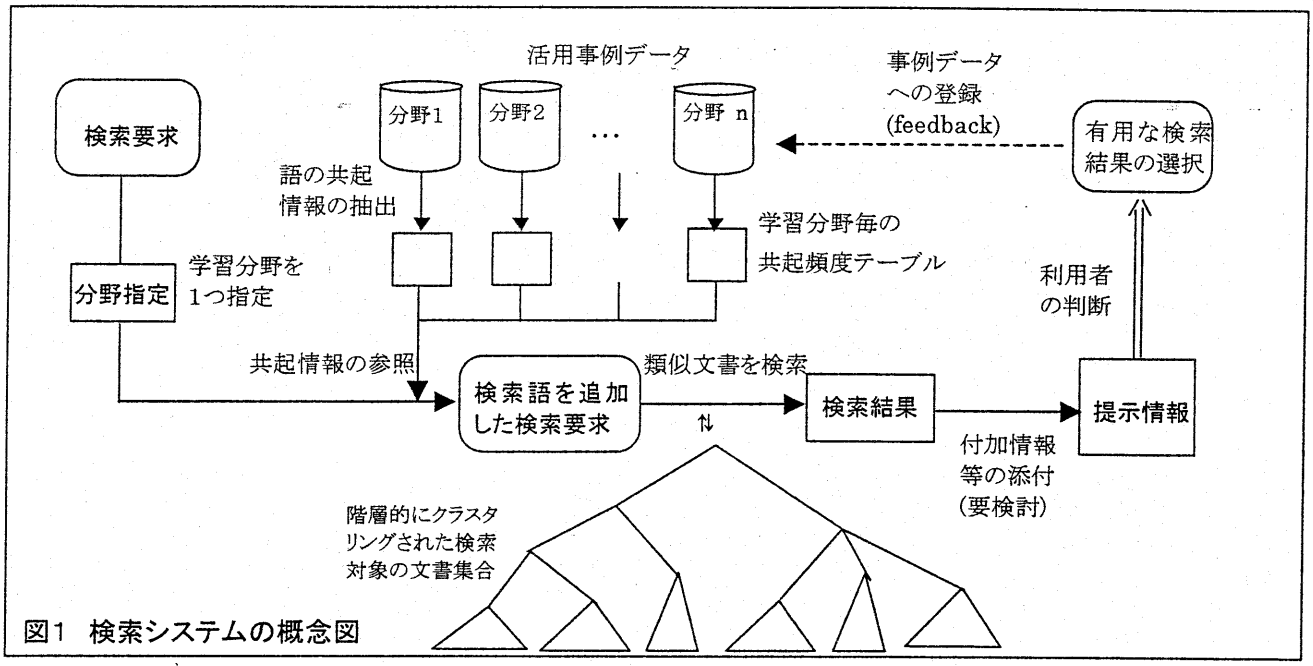

図1 検索システムの概念図

ージにおける文書の記述内容から，各分野毎の語の共起頻度(名詞のみに着目したもの)を抽出した。各ページの文書からの語の切り出しは，形態素解析ツール Chasen 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カウントした。

4. 共起情報を用いた検索要求の拡張

前章で述べたように，該当する学習分野内の共起情報を参照することにより，検索意図の反映に近い検索語の追加が行なわれると考えられる。ここで，入力される検索語が1語の単純な場合について，共起頻度の高い語を4語まで追加する例を表2に示す。

実際に検索要求の拡張の有無による検索精度・再現率等を比較するため，まず新聞記事(1998 年分のうち約 2 万件)について，各学習分野毎の出現頻度の比較的高い語群から任意に検索語を選び，これに対応する正解集合として学習利用可能と判断される記事を用意した。

検索実験においては，1 語の検索語に対して該当分野での共起頻度が高い語を順に追加して，検索結果を検討した。その結果，1〜3 語追加した場合に検索精度が向上する傾向が確認された。現状では，サンプル的な範囲ではあるが，学習分野内の代表的な共起語であれば，検索目的に対して寄与する確率が高いことが裏付けられた。

表2 共起頻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	彗星	＋発見，観測，日，月
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

5. まとめ

本稿では，学習情報源の語の共起情報を基に，検索語の追加を行う手法を検討し，理科における追加可能な検索語を抽出した。検索実験の結果，検索精度の向上が確認された。性能向上への，検索語への重み付けなどによる改良は今後の課題である。

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] 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.

[2] 鈴木雅実・村松茂樹・松本一則,・井ノ上直己・橋本和夫:"類似文書クラスタリング手法による新聞記事分野コード推定実験", 情報処理学会第 61 回全国大会, 2000.

[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.