自然言語処理 117－19
（1997. 1. 21）

語彙的連鎖を利用した文書検索

望月 源, 本田 岳夫, 奥村 学
Email:{motizuki,honda,oku}@jaist.ac.jp
北陸先端科学技術大学院大学 情報科学研究科

[概要]
最近の文書検索では，クエリと関連の高い文書の一部分を取り出して類似度を計算する passage レベルの検索が注目されている．この passage をどのように定義するかが問題となる．本稿ではこの passage を決定するのに，語彙的連鎖の情報を利用する手法について述べる．また，情報検索システム評価用ベンチマークを使用した検索実験を行い，語彙的連鎖の passage への有効性について検討する．
[キーワード] passage，語彙的連鎖，文書検索，BMIR-J1

Document Retrieval using Lexical Chains

MOCHIZUKI Hajime,HONDA Takeo,OKUMURA Manabu
School of Information Science, Japan Advanced Institute of Science and Technology
(Tatsunokuchi Ishikawa 923-12 Japan)

Abstract

The passage level retrieval has been recieved much attentions in the recent document retrieval task by researches, because a passage can be considered as a useful cue for extracting some effective part of a text. However it is still a problem how to define passages. In this paper we describe a definition of a passage using leixcal chains which denote local document contexts. We also present preliminary experiments of our retrieval method using BMIR-J1 which is bench mark for IR system.

Key Words passage, lexical chains, document retrieval, BMIR-J1

1 はじめに

文書検索は，入力されたクエリと文書集合内の各文書との類似度を計算することで的確な文書を選び出す作業である．またこの類似度に基づいて検索結果をランク付けすることが一般的である．

現在最も良く利用されている検索手法の1つにベクトル空間モデルがある [5]．ベクトル空間モデルでは，この類似度を計算するために，文書内の各タームに対して，文書内出現頻度 (tf)，文書集合での分散 (idf) に基づく重みをつけるなどの統計的な情報を利用する．しかし，検索対象の文章がある程度の長さを持つ場合，含まれるタームの数も多くなるので，単純な出現頻度だけでは統計量的に区別がつかない場合が起こり得る．例えば，図1のような場合，Doc.A と Doc.B のどちらが重要かという区別がつかない．特に長い文章の中には，たくさんの話題が含まれている．そのため，クエリに適合する文書の検索には，文書全体との比較だけよりも，文書内でクエリに関連する話題が述べられている部分を選び出し，その部分との類似度も計算した方が，より文書内容を反映した類似度を得られると考えられる．こうした点から，最近の研究では，文章を一定の長さに区切って扱う passage を用いた検索が注目されている [1, 3]．しかし，この場合どのように passage を定めるかという新たな問題が発生する．passage を考える場合，少なくとも何らかの方法で文章中の意味を捉え，クエリと最も関連のある部分を選び出す必要がある．

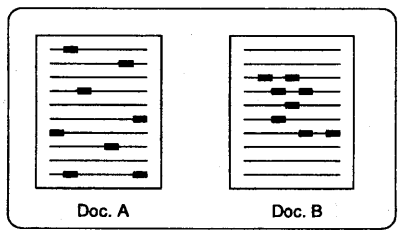

図 1: document の例

一方で，我々が文書中に出現する関連した語の連続を捉えるために利用できる情報として，語彙的結束性 [2] に基づく語彙的連鎖 [4] がある．語彙的連鎖は，文書中でその連鎖を形成する語彙に関する話題が述べられている範囲を表すものと考えられる．そこで，本稿では，我々は文書中の語彙的連鎖を利用して，クエリと対象文書の適切な関連度を計算できるような passage を提案する．またこの passage を利用し，検索結果のランク付け精度を向上させる手法の検討を行う．我々の手法の有効性を評価するため，計算機上に実装し，『情報検索システム評価用ベンチマーク BMIR-J1』[9] を用いて実験を行う．

2 passage レベルの文書検索

一般的に，passage を用いる検索では，文章中の各 passage とクエリの類似度を計算し，全てを足し合わせた値や最も高い値を passage による類似度とする．各文書とクエリの類似度は，passage による類似度と文書全体での類似度とを組み合わせて計算される．passage の決定に必要なことは，文書中でクエリに関連する部分を取り出すこと，クエリや文書に応じて柔軟にサイズが設定すること，文書内容に応じた有意なサイズを設定すること，であると考えられる．

現在の passage の単位は大きく分けて，文章の形式段落に基づくもの，意味段落に基づくもの，固定長 window に基づくもの，の3種類が考えられる．現実の文章では同一の話題が複数の段落に跨る場合もある．また，形式段落の切れ目と内容の切れ目が一致しない場合がある [7]．そのため，形式段落を passage の単位とすると検索精度に悪影響を与える可能性がある．また，文書を固定長 window に区切った passage を単位にすると，有意な window サイズを決定しなければならないという問題が生じる．実際には，効果的な検索を実現するために，データベース中の文書の長さや種類によって調整する方法がよくとられている．しかしこの方法では，passage の決定に文書の意味的な要素が反映されていない．また，固定長の passage が実際の文章の話題の範囲に一致しているとは考え難いという問題がある．

本来文書中で特にクエリと関連する部分を取り出して類似度を計算することが，passage を採用する理由であるから，passage を決定する場合に，文書中の意味をとる努力は必要である．そのため意味段落に基づく passage は，文書内容に基づいて計算される単位な

ので，最も望ましい方法だと思われる．しかし現在のところ，厳密な意味段落への分割は期待できない[6]．そこで，今回我々は厳密に意味段落を計算する代わりに，比較的浅い方法を用いて文書の意味を取り出すことにした．比較的浅い意味解析としては，語彙的連鎖が利用できる．語彙的連鎖はテキスト中に出現する意味的に関連した語の連続である．一般に文書には複数の語彙的連鎖が存在し，文書構造の結束性を表している．つまり，ある語彙に関する連鎖が出現する部分では，その語彙に関する話題が述べられていると考えることが出来る．この語彙が入力されたクエリと一致する場合を考えると，文書中に現れる語彙的連鎖の範囲が，クエリに強く関連する部分であると見ることが出来る．この語彙的連鎖はシソーラス上で同一カテゴリにある語の連続として計算できる．

そこで我々は，passageを次のように考える．入力されたクエリ内の各クエリタームを含む語彙的連鎖を部分passageと見立る．passage全体は，各部分passageを組み合わせたものとする．クエリと文書の類似度も，各部分passageごとの類似度を組み合わせることで計算する．

この手法により，クエリと意味的に関連する部分を取り出し，文書内容を反映したpassageを設定することが期待できる．また，ある程度文章内容を反映する連鎖のみを連鎖と認めるという制約をかすことで，意味あるサイズを持ったpassageを決定できる．さらに，シソーラスを用いて連鎖を計算することで，クエリのシソーラス展開と同様の効果も得られる．

Callan[1]が，固定長window passageを改良した検索手法を提案している．Callanの手法では，入力されたクエリと最初に一致した場所からwindowを開始し，以降のwindowは前のwindowの中間点から開始される．これは単なる固定長windowに比べてクエリに関連する部分を取り出す工夫がされている．しかし，必ずしもクエリの最初に出現する位置から話題が開始されるとは限らない．一方，我々の手法では，クエリを含む連鎖を形成する語の開始点を取り出せる．また，Callanの手法では，windowを用いるpassageの問題点は以前として残っている．

3 実装

我々の検索は文書全体での検索とpassageレベルの検索の2つにわかれる．そのため異なるインデックスをそれぞれ用意する．入力クエリに対する検索は伝統的なベクトル空間法を使用する．これにより得られた2レベルの類似度を組み合わせることにより最終的な各文書との類似度を求め，値の高い順に利用者に提示する．

3.1 文書全体

文書全体での検索では，各文書を次のようにインデクシングする．まず，文書を形態素解析し，名詞，動詞，形容詞をタームとして選択する．次に、各タームに式(1)の$tf \cdot idf$法によって重み付けを行う．

$$w_{ik} = \frac{tf_{ik} \times log(N/n_k)}{\sqrt{\sum_{j=1}^{t}(tf_{ij})^2 \times (log(N/n_j))^2}} \quad (1)$$

ただし，w_{ik}は文書D_iに存在するタームT_k，tf_{ik}はD_iのタームT_kの出現頻度，Nは文書の数，n_kはT_kを含む文書の数であり，分母は，文書の長さによる正規化を意味する．

このようにして作られたインデックスが与えられて，実際の検索は次のように行う．

まず入力されたクエリをインデクシングと同様に形態素解析を行いクエリタームを選択する．次にクエリタームを文書と同様にベクトル表現する(Q_j)．次にクエリベクトルQ_jと文書ベクトルD_iの類似度$sim(Q_j, D_i)$を式(2)の内積によって求める．

$$sim(Q_j, D_i) = \sum_{k=1}^{t} w_{jk} w_{ik} \quad (2)$$

類似度の高い文書ほどクエリの内容を反映していると考える．

3.2 passageレベル

passageレベルの検索では，各文書を次のようにインデクシングする．

まず，文書を形態素解析し，名詞，動詞，形容詞を選び出す．次に，我々のシソーラスを用いた語彙的連鎖計算アルゴリズム[11]により文書内の語彙的連鎖を

計算する．この段階で，図2のようにいくつかの同一概念を持つ単語の連鎖が生成される．しかし例えば，図2の連鎖Cでは，連鎖を形成する単語が極端に少ないので，有意な連鎖と認めない方が良いと思われる．また連鎖Aでは，8語目から17語目までの間，単語が出現しない我々がギャップと呼ぶ部分が存在する．この場合，ギャップで連鎖を切り離して，別の連鎖として計算した方が良いと思われる．

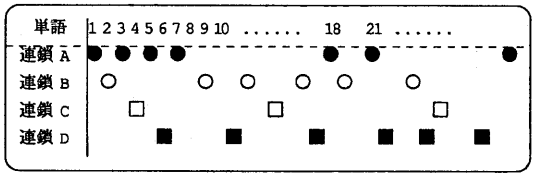

図 2: 語彙的連鎖のイメージ

そこで次のステップとして，有意な連鎖を選ぶために，ギャップや連鎖の長さを考慮した以下の制約をかける．

- ギャップ長閾値を設定し閾値以上出現しない場合はそこで連鎖を切る
- 連鎖長閾値を設定し，閾値以上の長さを持つ範囲だけを連鎖とする

図2の連鎖Aを例に考えると，あるギャップ長閾値のために，連鎖Aはそれぞれ部分連鎖A1，A2，A3，に分割される．次にそれぞれの部分連鎖に対して連鎖長閾値の制約をかけることで，部分連鎖A2，A3は，連鎖と認められなくなる．結果として，元の連鎖Aの中で，部分連鎖A1だけが有意な連鎖として残される(図3)．

この制約により残ったものだけを連鎖と認める．

インデクシングタームは連鎖を形成するシソーラス上の概念番号になる．また，シソーラスの見出し語になっていない語で文書内に繰り返し出現する語についても連鎖と同様の制約を満たすものについては，タームに含めている．従ってインデックスタームは，シソーラス上の概念番号と未知語が混在した形をしている．なお，今回はシソーラスに角川類語新辞典 [8][1]を使用する．

[1]概念番号の上位3桁を同一概念とする

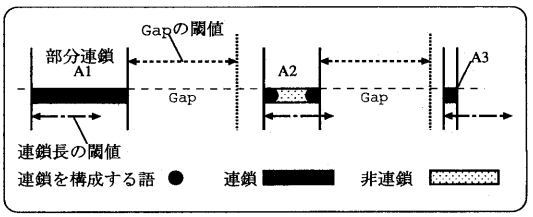

図 3: 連鎖の決定

また全体での検索と同様に，各タームには式(1)により重みをつける．ただしこの場合，式(1)のw_{ik}は連鎖D_iに存在するタームT_k，tf_{ik}はD_iのタームT_kの出現頻度，Nは文書の数，n_kはT_kを含む文書の数であり，分母は，連鎖の長さによる正規化を意味する．

有意な連鎖を認める時の制約により，タームT_kを含む文書であっても，そのT_kが有意な連鎖を形成していない場合は，idfの計算に含まれていない点に注意して欲しい．例えば図2の文書中に出現する連鎖Cが有意な連鎖と認められなかった場合に，この文書は$\text{ターム}\ T_{\text{連鎖}\ C}$を含むことにならない．つまりこの場合の$idf$は『特定の文書に意味ある出現をするタームに高い重みをかける』と解釈できる．

上記の重みつきインデックスが与えられて実際の検索は次のようになる．

まず入力されたクエリをインデクシングと同様に形態素解析を行いクエリタームを選択する．次に，シソーラスを索いて，複数の概念番号を得る(未知語の場合は，そのままにする)．各クエリの各概念ごとにマッチする文書とその文書でのスコアを計算する．同じクエリから得られた複数の概念が同一文書でマッチする場合には，値の大きい方を優先する．それぞれのクエリについて計算し，式(3)で，全てのクエリについての類似度を足した値をその文書とクエリの類似度とする．

$$sim(Q_j, D_i) = \sum_{k=1}^{t} \max_{l=1\cdots c} (w_{jkl} \times w_{ikl}) \qquad (3)$$

ただし，lはタームT_kの持つシソーラス上の概念を意味する

厳密には各クエリの全ての概念の組合せで類似度を計算することになるが，検索時間にかかるコストを軽減するためにこのように近似する．

3.3 全体と passage レベルの組合せ

2つのレベルの類似度の組み合わせについては，様々な方法が考えられるが，今回は最も単純な手法を用いる．式(4)で表すように，それぞれ最大の類似スコアで割った値を足すことで求める．

$$sim(Q_j, D_i) = \frac{sim_{all}(Q_j, D_i)}{max(sim_{all})} + \frac{sim_{pas}(Q_j, D_i)}{max(sim_{pas})} \quad (4)$$

ただし，$sim(Q_j, D_i)$ は，2つのレベルを組み合わせた類似度，sim_{all} は文書全体での類似度，sim_{pas} は passage レベルでの類似度であり，max はそれぞれの類似度の最大値を表す．

4 評価実験と考察

4.1 実験

我々の手法の有効性を評価するためシステムを実装し，実験を行った．

実験には，情報検索システム評価用ベンチマーク BMIR-J1 ver.1.0[9] を使用した．BMIR-J1 は，検索対象テキスト集合 600 件 (日本経済新聞の朝刊経済面記事)，クエリ 60 種とそれに対する正解がセットとなっているベンチマークである．また，BMIR-J1 ではクエリに5種類のファンクションが対応付けされており，研究課題に応じて使用するクエリと正解の組を選ぶことが出来るようになっている．この内，今回は以下の2種類のクエリと正解のセットを使用した．ただし，我々の今回の実装は，BMIR-J1 で想定している機能の内，基本機能に対応しているので，正確に評価できるのは1のセットである．2のセットについては，1のセットのクエリが 10 と少ないため，参考のために行った．

1. 基本機能 (F1)[2]だけを必要とする 10 のクエリと正解のセット
2. 数値レンジ機能 (F2)[3] を必要としない 55 のクエリと正解のセット (1 を含む)

上記の2セットを用いて以下の6つの組合せで，実際に文書集合 600 件に対して，検索を行った．

1. セット 1 を用い，文書全体での類似度のみで計算
2. セット 1 を用い，passage での類似度のみで計算
3. セット 1 を用い，passage と文書全体での類似度を組み合わせて計算
4. セット 2 を用い，文書全体での類似度のみで計算
5. セット 2 を用い，passage での類似度のみで計算
6. セット 2 を用い，passage と文書全体での類似度を組み合わせて計算

なお，今回は語彙的連鎖のギャップ長閾値，連鎖長閾値はそれぞれ各文書のインデックスタームの $\frac{1}{4}, \frac{1}{12}$ とした．

評価は，検索された文書について，2つの評価尺度である，式(5)の recall と式(6)の precision を用いて行う．なお，BMIR-J1 では，正解文書がその判定基準により A,B,C にわかれているが，今回は A と B を正解文書として評価した．

$$recall = \frac{\text{システム出力の正解文書数}}{\text{ベンチマークの正解文書数}} \quad (5)$$

$$precision = \frac{\text{システム出力の正解文書数}}{\text{システムの出力文書数}} \quad (6)$$

実験1から3までの結果の平均を図4に，実験4から6までの結果の平均を図5にそれぞれ示す．また，セット1を用いる実験1から3の中で代表的な特徴を示した3つのクエリについて図6, 図7, 図8に示す．

4.2 考察

全クエリ平均で見ると，実験セット1と2では，全体的にセット1での結果の方が上回っていた．これは，もともとセット1の方を想定していたため，当然の結果だと考えられる．また，どちらの場合も，passage での類似度のみ (図 4の Ex2 と図 5の Ex4) の時が最高の結果だった．

[2]基本機能とは,『キーワードの存在確認，あるいはキーワードのシソーラスによる展開語の存在確認．および，それらの語の存在に関する論理式 (AND や OR など) の充足判定など.』[9] を意味する

[3]数値レンジ機能とは,『数の数え上げや，数値などの範囲を正しく解釈する．数値の大小比較や単位の理解・変換なども含む.』[9] を意味する．他の機能は，構文解析機能 (F3)，内容解析機能 (F4)，知識処理機能 (F5) である

セット1の10クエリについて分析すると，文書全体での検索とpassageのみの検索との間に次のような場合が見られた．

1. どちらの単位によっても検索できなかったもの
   10クエリの内，1つがこれに該当した．『ビデオデッキ』というクエリに対して，定められた正解文書中では「ＶＴＲ」がマッチしなければならないものであった．しかし，今回我々が用いたシソーラスではＶＴＲという見出し語は用意されていない．そのため正解を検索することが出来なかった．

2. 文書全体に比べて，passageでの検索結果の方がprecisionが向上するもの (例:図6)
   これは5クエリに見られた．この場合，語彙的連鎖のpassageによる類似度が良い結果に継ったと考えられる．例として,『国内航空大手３社』を入力とする場合を考える．これは形態素解析により『国内，航空，大手，３，社』の5つのクエリタームに分割される．文書全体での類似度を見ると，1つか2つの高い重みを持ったタームにマッチする文書が上位にきて不正解となることが多かった．一方，passageでは2つか3つ以上のタームが高い類似度でマッチする文書ほど上位に来ており，正解となることが多かった．このことからも，文書全体と類似度を計算するよりも，語彙的連鎖によるpassageによって計算する方が適切なスコアが計算されたと考えられる．

   『航空(概念番号316)』を含む語彙的連鎖により取り出された部分passageの実例を付録に示す．

3. 逆に，passageによって文書全体に比べてprecisionが下がったもの (例：図8)
   これは2クエリに見られた．この場合考えられる原因として，入力クエリが一般的な語であることが考えられる．例えば『減税』は，角川類語新辞典において概念番号262に一意に決まるが，同概念の単語が103語ある．これは1概念当たりの平均単語数55.77語[10]に比べて非常に多い．実際，今回の文書集合中，同一概念として連鎖を形成する他の単語には,『減，減少，縮小，削減，増，増加』などがあり,『減税』よりも多数出現していた．一方文書全体の検索では，文字列一致であるために，同概念の語による問題は起こらない．

4. 文書全体での類似度を用いた検索では，recallが1にならなかったが，passageのみでは再現することが出来たもの (例：図7)
   2クエリに見られた．これはクエリと同じ文字列が出現しない正解文書の存在を，我々の語彙的連鎖に基づくpassageによって概念レベルで検索できた例である．クエリ展開と同様な効果を引き出すことを示している．

考察2から，語彙的連鎖によるpassageの利用によって，文書中でクエリと関連の強い部分を取り出せる可能性があることがわかった．一方考察3に見られるように，クエリによっては，逆にprecisionを下げる可能性もある．このような場合には全体での検索の方が良い結果を得られる．一方考察4から語彙的連鎖によるpassageのインデクシングがクエリ展開と同様の効果も期待できる．このことから，より正確で強力な検索を実現するためには，2つの類似度をうまく組み合わせて類似度を計算することが必要である．この点について，今回我々は簡単な組合せを実験3,6として行ったが，良い結果を得られなかった．

5 おわりに

今回の予備的な実験から，語彙的連鎖によるpassageが，文書検索精度の向上に有効であることがわかった．また，インデクシング段階で，語彙的連鎖を使用することにより，クエリ展開と同様な効果を得られることもわかった．これらは語彙的連鎖の利点である．しかし一方で，概念でのマッチングのためにprecisionが下がる場合も見られた．このような時には文書全体での類似度の方が良い結果が出た．そのため，2つのタイプの類似度を組合せる実験も行ったが，今回の組合せ手法では両方の良い点を出すことができなかった．

今後の課題としてまず，組合せのチューニングがあげられる．この他にも，ギャップ長閾値，連鎖長閾値，部分passageによる類似度の組合せ方法では，パラメータチューニングが行える．これらのチューニングについても試してみる必要がある．

今回はベクトル間類似度に内積を用いている．これは論理型検索の OR 検索に近いと考えられる．しかし，実際には入力されたクエリ間には，AND 的な結合や OR 的な結合があると思われる．そこで，その結合タイプを推定し，それぞれ重みを変える，制約を加えるなどの点も考慮していく．また，考察で述べたように，*tf.idf* のような出現頻度による重み付け要因以外に，各単語のシソーラス上での一般性が，検索精度に影響を与えることがわかった．そこでこの観点からの重み付けについても検討していく．

謝辞

本研究でのベンチマーク「BMIR-J1」の使用許可を下さった，情報処理学会データベースシステム研究会，情報検索システム評価用データベース構築ワーキンググループに感謝致します．また本研究で使用した，「角川類語新辞典」の使用許可を下さった (株) 角川書店に感謝します．

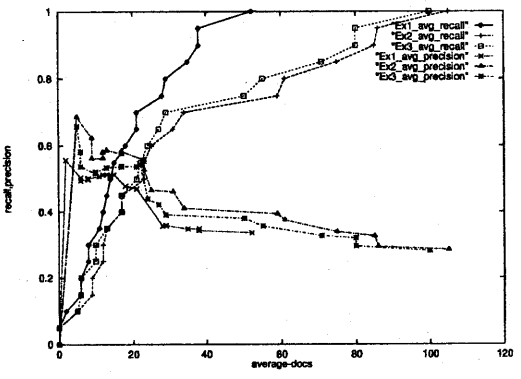

図 4: 実験 1,2,3 の平均

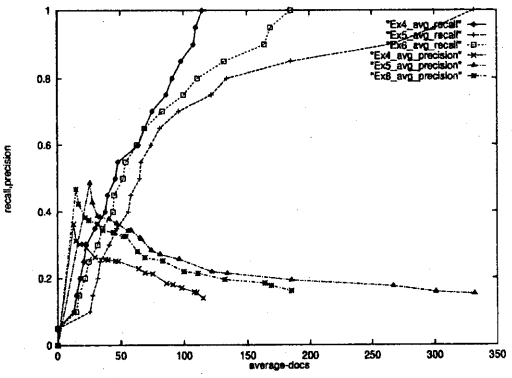

図 5: 実験 4,5,6 の平均

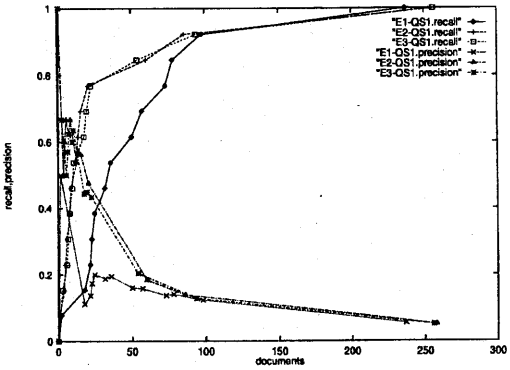

図 6: クエリ『国内航空大手３社』の結果

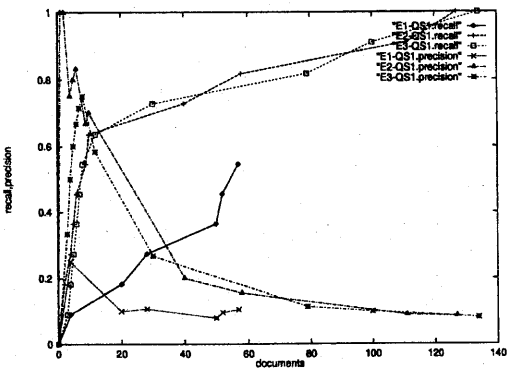

図 7: クエリ『飲料品』の結果

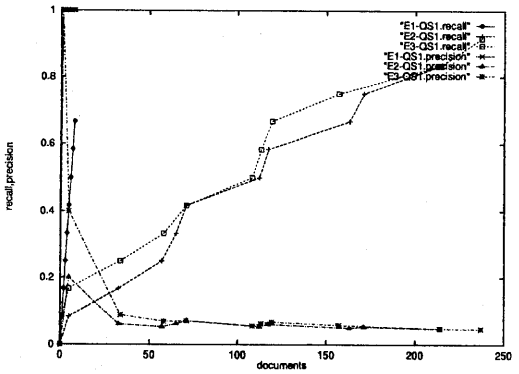

図 8: クエリ『減税』の結果

参考文献

[1] J. P. Callan. Passage-Level Evidence in Document Retrieval. In *Proc. of 17th Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval*, pp. 302–310, 1994.

[2] H.A.K. Halliday and R. Hasan. *Cohesion in English.* Longman, 1976.

[3] M.A. Hearst and C. Plaunt. Subtopic Structuring for Full-Length Document Access. In *Proc. of 16th Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval*, pp. 59–68, 1993.

[4] G. Morris, J.and Hirst. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, Vol. 17, No. 1, pp. 21–48, 1991.

[5] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley, 1988.

[6] 望月源, 本田岳夫, 奥村学. 重回帰分析とクラスタ分析を用いたテキストセグメンテーション. 第 2 回 言語処理学会年次大会 C2-5, pp. 325–328, 1996.

[7] 所一哉. 現代文レトリック読解法. 匠出版, 1987.

[8] 大野晋, 浜西正人. 角川類語新辞典. 角川書店, 1981.

[9] 福島俊一, 小川泰嗣, 石川徹也, 増永良文, 木本晴夫, 田中智博, 中渡瀬秀一, 芥子育雄, 豊浦潤, 宮内忠信, 上田良寛, 松井くにお, 三池誠司, 酒井哲也, 木谷強, 徳永健伸, 鶴岡弘, 安形輝. 日本語情報検索システム評価用テキストコレクション BMIR-J1. 自然言語シンポジウム「大規模資源と自然言語処理」, pp. 1–16, 1996.

[10] 望月源, 本田岳夫, 奥村学. 語彙的結束性に基づいた語義曖昧性解消の観点から見たシソーラスの比較. 情報処理学会第 51 回全国大会論文集 3H-2, pp. 37–38, 1995.

[11] 本田岳夫, 奥村学. 語義曖昧性を考慮した有意な語彙連鎖の生成. 情報処理学会研究会資料 NL97-14, pp. 95–102, 1993.

付録

『航空』(概念番号 316) に強くマッチする部分 passage の例:

『航空など国内航空会社は景気の低迷と国際競争の激化で業績が悪化、九二年度は日航が五百三十八億円、日本エアシステムが四十八億円の経常赤字を出す苦境に陥っている。航空』

『航空会社による違いもほとんどないのが実情で、顧客管理の手法を駆使して多様な割引制度を開発した米国の航空会社とは対照的だ。国内の航空市場は主要航空三社の寡占状態』

『航空は二十九日、九月中間決算を発表、経常赤字は七十九億円に拡大した。日本エアシステム（ＪＡＳ）も赤字になり、全日本空輸も大幅減益になった。長引く景気低迷による収入減に加え、高コスト体質からの脱却が遅れているためだ。すでに設備投資の抑制や経費の節減などの合理化策に乗り出しているものの、効果は十分上がっていない。収益改善のための国内航空運賃の値上げ論議も出ているが、スンナリ値上げは通りそうにない。航空業界は、一層厳しい合理化が迫られている。航空三社の九月中間決算は航空不況の深刻化を一段と浮き彫りにした。日本航空』