社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－DBS－137 (II) (36)
2005／7／14

検索語の曖昧性を解消するキーワードの提示手法

若木 裕美† 正田 備也‡ 高須 淳宏‡ 安達 淳‡

†東京大学 情報理工学系研究科
‡国立情報学研究所

既存の検索エンジンでは，キーワードを用いた検索が主流であり，それゆえ数語による適切な検索質問の組み合わせを見つけるのが容易ではない．本稿では，単語の低頻度共起を利用して，検索質問の曖昧性を解消する手法を提案する．これは，"いろいろな種類の単語と共起する単語は，独立したトピックを持つことができない" という仮説に基づき，単語に重みを与えることである．

本手法によって見出された単語を，元の検索質問に追加すると，平均適合率の上昇に大変効果があることを示す．また，他の手法に基づく上位の語に比べると，より細かく特定の内容を示すグループに分けるように作用する語である．検索質問に合致するような単語を見出すことを可能とする．

Query Ambiguity Indication Using Infrequent Term Cooccurrences

Hiromi WAKAKI†, Tomonari MASADA‡, Atsuhiro TAKASU†, and Jun ADACHI‡

†Graduate School of Information Science and Technology, The University of Tokyo
‡The National Institute of Informatics

Conventional search engines are designed mainly for general keyword search. Therefore, in many cases, we can find no appropriate combination of query terms. In this paper, we present a query disambiguation method by using infrequent term cooccurrences. This strategy comes from the following idea: terms appearing with a wide variety of terms cannot establish an independent topic. Based on this hypothesis, terms are weighted.

The experimental results show that the terms ranked higher by our method can improve the average precision of Web search when added to the original query terms. As compared with other term ranking methods, our method gives higher ranks to the terms denoting more particular and adequate stuff and referring to more specific concepts.

1 はじめに

電子的情報の爆発的増大は，従来型の検索エンジンによる調べ物を困難にしている．ランキングされた結果の出力を見るだけでは所望の情報が得られず，無駄に使う時間が大きくなっている．

欲しい結果が正しく得られにくいと感じた時，検索エンジンが悪かったのであろうか．ユーザによって入力される検索語は信頼性の高いものと言えるだろうか．Web 上にあるぼう大な情報は，元々体系的に作られたものではない．現状では，質問者が欲しい情報を得るために，Web 上の情報の特性に合わせて「Web 上に存在するデータから，的確に欲しい情報を持ってこれる語」を入力する必要がある．また，さらに質問語が 1，2 語であることが通例であり，圧倒的に質問の持つ情報量が少ないことも質問処理を困難にしている．

例えば，google を使って「スキー」を検索語として結果を見てみると，上の方にリストされるのは「ス

キー場」のオフィシャルなページである．そこで，「スキー －場」(「スキー」∧「¬場」の意味) と入力して結果をみると，連盟や協会，部等の団体のページが上に多くあがってくる．スキーから思い浮かべるものは人それぞれであるが，その結果が「スキー場」や「スキー連盟」に押しつぶされてしまう．多様なものを，1つのものさしで測った結果の上位しか見ることができない結果になっている．また，曖昧な状態で検索すれば，キーワードを見つけるために色々な語を試して的確な言葉を探すことを要す．

本稿では，検索語をヒントにユーザの必要とする語を提示する手法を提案する．すなわち，『検索語からみるとより細かいトピックを示す語（少しマイナーな語）であるが，検索語から連想するものとして的を得ている語』（以下，これを "Articulateness" と呼ぶことにする．）を集める (図 1)．将来的には，その中でまとまったトピックに分けて提示することで，ユーザは自分の検索要求の中にある曖昧性に気が付き，幾つかのトピックの中からより自分の要求に合致するものを選択することが可能となる手法であると考えている (図 1 右)．

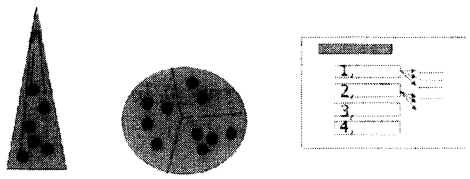

図 1：本研究のねらい．(1，主要なトピックに押しつぶされてしまうトピックに含まれる単語も取り込む．2，トピックごとに単語を分ける．3，ユーザが一見して理解しやすいような提示を行う．)

2 Articulateness(分節性)

2.1 基本的な考え方

2.1.1 前提

語と語の共起が頻繁に起こる語同士は，概念的に近い（連想されやすい）語であると考える (図 2)．なお，同じ文書に現れることを共起と定義する（図 3)．

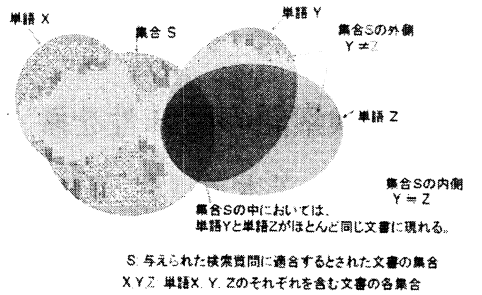

図 2：ある文書集合の中における共起の意味．

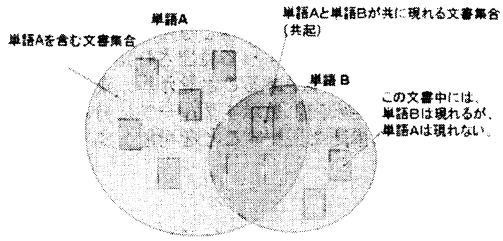

図 3：共起の定義

2.1.2 分節性の仮説

検索質問に含まれる曖昧性に対し，検索から得られる結果をトピックに分けて提示することを想定する．既存の特徴語抽出では，検索で得られる文書集合の多くの文書に現れる語のスコアが高いため，その文書集合の中でトピックに分けることは難しい．そこで，元の文書集合の中に含まれる幾つかのトピックに分かれるように語を選ぶように，既存の方法とは異なる指標で，適度にマイナーな語を集めてくるアルゴリズムを考える．ここで，各トピックについて特徴的である単語ほど（1つのトピックには現れるがそれ以外のトピックには現れにくいほど），『Articulateness(分節性) が高い』ということにする．

このような理想に合うモデルとして，**「その文書集合の中において，多くの種類の語と一緒に出てくる可能性の高い語は，その文書集合の中で独立したトピックをもてないものである」**と仮定する．

ここでは，分節性を評価する指標として，以下に AR1 と AR2 を提案する．そして，既存の様々な指標との相互比較をしてみる．

2.2 AR1:異なり語数に基づく分節性の定式化

ある程度多くの文書に登場する語（マイナー過ぎないようにする意図）であるが，その語と**一緒に出てくる語の種類が少ないもの（独立したトピックをもてる語）**が，スコアが高くなるように定式化を行う．まず，次のように定義する．

S：検索質問に適合した文書集合
U：検索の対象となる文書全部の集合
$N_S(t_i)$：Sの中で，t_iが含まれる文書の数
$N_U(t_i)$：Uの中で，t_iが含まれる文書の数
N_S：集合Sに含まれる文書数

ここで，単語t_iと一緒に出てくる語の種類は，次の式で表せる．

$$\sum_{t_j} \frac{N_S(t_i \wedge t_j)}{N_S(t_i)} \tag{1}$$

これは，t_iが現れる文書の平均異なり語数であるので，$AvgType(t_i)$と書くことにする．$AvgType(t_i)$が小さい語ほど，独立したトピックを持てる語と考えられ，分節性の指標として使うことができる．しかし，$N_S(t_i)$が小さい単語は，分節性に関係なく，$AvgType(t_i)$も小さくなりやすい．そこで，更に，ある程度以上多くの文書に登場する語であるという条件を加えて，

$$\begin{aligned} AR1(t_i) &= \frac{N_S(t_i)^2}{N_U(t_i)} \times \frac{1}{AvgType(ti)} \\ &= \frac{N_S(t_i)^2}{N_U(t_i)} \frac{1}{\sum_{t_j} \frac{N_S(t_i \wedge t_j)}{N_S(t_i)}} \end{aligned} \tag{2}$$

という特徴量で単語t_iの文節性を表すことにする．$\frac{N_S(t_i)^2}{N_U(t_i)}$という項は，$\frac{N_S(t_i)}{N_U(t_i)}$と$N_S(t_i)$を掛け合わせてつくった項である．$\frac{N_S(t_i)}{N_U(t_i)}$は，全体$U$の中での現れ方に比べて，集合$S$の中での方がどれくらい現れやすくなったかという割合であり，単語t_iが集合Sにどれくらい関係しているかを相対的に表している．$N_S(t_i)$は，集合Sの中での出現頻度そのものであり，単語t_iが集合Sにどれくらい関係しているかを絶対的に表している．そして，この項を平均異なり語数で割ることにより，「一緒に出てくる語の種類が小さい語は分節性が高い」という分節性の仮定を，定式化することができる．

2.3 AR2:確率論的観点からの分節性の定式化

もう1つの定式化として，集合Sでの共起頻度と，単語t_iを含む文書集合の中での共起頻度とのずれを計算することで，単語t_iの分節性を評価する式を考える．出現頻度に関する単語同士の影響度を測るために，情報量的な式を導入する．

2.3.1 単語の出現確率と共起確率

「単語t_iが文書に現れる」確率を$P_S(t_i)$，ひとつの文書中で単語t_iとt_jが共起する確率を，$P_S(t_i \wedge t_j)$と書くことにする．本稿では，

$$P_S(t_i) = \frac{N_S(t_i)}{N_S}$$
$$P_S(t_i \wedge t_j) = \frac{N_S(t_i \wedge t_j)}{N_S}$$

と定義する．以下，出現確率，共起確率は，集合S上で考えるので，添え字Sは省略し，$P(t_i)$，$P(t_i, t_j)$などと書く．

2.3.2 相互情報量と分節性の違い

このように確率を設定し，相互情報量(MI)を計算すると(以下の式(6)にあるように)，

- 単語t_iが現れていることに対する，単語t_jの関連度
- 単語t_iが現れていないことに対する，単語t_jの関連度

の二つを混合した値となる（図4）．

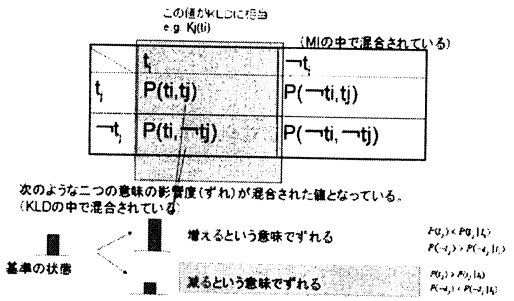

図4：MIとKLDの違いと，KLDと分節性の違い．

ここで，

$$K_j(x) = \sum_{y \in t_j, \neg t_j} P(y|x) \log \frac{P(y|x)}{P(y)} \tag{3}$$

と書くことにすると，相互情報量の式は

$$I(X:Y) = P(t_i)K_j(t_i) + P(\neg t_i)K_j(\neg t_i)$$

と変形できる．つまり，相互情報量は，$K_j(X)$ の期待値となる．この $K_j(X)$ は，Kullback-Leibler divergence と呼ばれる値である．

分節性を考える上で関心のある値は，『単語 t_i が現れている文書での単語 t_j の現れ方に対して，文書集合 S での単語 t_j の現れ方がどれくらいずれているか．』であり，これは $K_j(t_i)$ が大きくなる場合だけである．

2.3.3 KL 情報量と分節性の違い

$K_j(t_i)$ の値が大きくなる場合というのには，次の二つの場合が考えられる．

(a) $P(t_j|t_i)\log\frac{P(t_j|t_i)}{P(t_j)} < 0$，$P(\neg t_j|t_i)\frac{P(\neg t_j|t_i)}{P(\neg t_j)} > 0$ となることによって，$K_j(t_i)$ が大きくなる場合

(b) $P(t_j|t_i)\log\frac{P(t_j|t_i)}{P(t_j)} > 0$，$P(\neg t_j|t_i)\frac{P(\neg t_j|t_i)}{P(\neg t_j)} < 0$ となることによって，$K_j(t_i)$ が大きくなる場合

KL 情報量では，この両方の項目が同時に考慮されている（図 4）．つまり，単語 t_i が出現する文書に限ったときに，他の単語の出現確率が，**増えるのであれ，減るのであれ，どれぐらい大きくずれるか**を調べる時に使うのが，KL 情報量である．

しかし，本稿で扱いたい単語の分節性の指標としては，**他の単語の出現確率が低くなるかどうか**，だけが問題であり，前者の項 (a) だけが必要となる．

2.3.4 分節性の第二の定式化

分節性の考え方を情報量的な式に合わせると，**『他の単語の出現確率が低くなる方向にずれるものが良い』**というモデル化を得る．

この意味での分節性は，KL 情報量の式を分解して定式化することができる．つまり，$P(t_j|t_i)\log\frac{P(t_j|t_i)}{P(t_j)}$ がマイナスになり，かつ，$P(\neg t_j|t_i)\log\frac{P(\neg t_j|t_i)}{P(\neg t_j)}$ がプラスになるような単語 t_j が多いほど，『単語 t_i の分節性は高い』と定義でき，AR2 で表すことにする．ただし，『他の単語の出現確率が低くなる方向にずれる』ことを次の SKL で表すことにする．

$$SKL(t_j;t_i) = -P(t_j|t_i)\log\frac{P(t_j|t_i)}{P(t_j)} + P(\neg t_j|t_i)\log\frac{P(\neg t_j|t_i)}{P(\neg t_j)}$$

$$AR2(t_i) = \frac{N_S(t_i)^2}{N_U(t_i)} \times \sum_{t_j \neq t_i} SKL(t_j;t_i)$$

と書ける．[1]

3 実験

3.1 実験における比較対照の式の整理

単語 t_i と他の各単語 t_j との共起情報に基づく特徴量を $w(t_i,t_j)$ とする．そして，

$$CW(t_i) = \sum_{t_j \neq t_i} cw(t_i,t_j) \tag{4}$$

と定義される $CW(t_i)$ を，t_i と他の単語の共起情報すべてを集約した値とする．さらに，この $CW(t_i)$ に $N_S(t_i) \times \frac{N_S(t_i)}{N_U(t_i)}$ をかけることで，文書集合 S に強く関係する単語ほど，共起に基づく特徴量がより強調されるように，単語のスコアを定める．

$$W(t_i) = N_S(t_i) \times \frac{N_S(t_i)}{N_U(t_i)} \times CW(t_i) \tag{5}$$

機械学習による文書分類における特徴語選択の代表的な方法として document frequency (DF), mutual information (MI), χ^2検定(CHI) などが挙げられる [9]．また，Query Expansion では特徴語選択として，RSV という指標も用いられる [5]．

今回の実験では，以下の 8 通りのスコア付け手法を比較した．

- AR1
- AR2
- $N_s(t_i) \times \frac{N_S(t_i)}{N_U(t_i)}$ の部分だけ (UnitWeight)
- CF
- 相互情報量 (MI)
- KL 情報量 (KL)
- χ^2 検定
- RSV

本稿で提案する AR1,AR2 に対して，比較する各々の式と，各式の持つ意味について以下に述べる．

[1] ここで新しく定義した SKL に近い式を，Lau らが [3] の論文において，異なる用途で使っている．

3.1.1 UnitWeight

$CW(t_i) = 1$とし，$N_S(t_i) \times \frac{N_S(t_i)}{N_U(t_i)}$の部分だけを単語のスコアとする．これによって，すべてのスコア付け手法に共通する項の影響を見ることができる．

3.1.2 CF

式(2)とは逆の考え方，すなわち，『他の語と共起する回数の**多い方**が良い語である』という指標を定式化すると以下のようになる．比較対照のため，この式についても実験を行った．

$$CF = \frac{N_S(t_i)^2}{N_U(t_i)} \times AvgType(t_i)$$
$$= \frac{N_S(t_i)^2}{N_U(t_i)} \times \sum_{t_j \neq t_i} \frac{N_S(t_i \wedge t_j)}{N_S(t_i)}$$

3.1.3 相互情報量 (MI)

語の共起に合わせると，『単語t_iの**有無**と，単語t_jの**有無**が，互いにどれだけ影響し合っているか.』という量を表しており，次のような式で書ける．

$$MI(t_i, t_j) = P(t_i)\{P(t_j|t_i) \log \frac{P(t_j|t_i)}{P(t_j)}$$
$$+ P(\neg t_j|t_i) \log \frac{P(\neg t_j|t_i)}{P(\neg t_j)}\}$$
$$+ P(\neg t_i)\{P(t_j|\neg t_i) \log \frac{P(t_j|\neg t_i)}{P(t_j)}$$
$$+ P(\neg t_j|\neg t_i) \log \frac{P(\neg t_j|\neg t_i)}{P(\neg t_j)}\} \quad (6)$$

$$W(t_i) = \frac{N_S(t_i)^2}{N_U(t_i)} \times \sum_{t_j \neq t_i} MI(t_i, t_j)$$

3.1.4 KL 情報量 (KL)

語の共起の書き方に置き換えると，『単語t_iの**有る**ことが，別の単語t_jの**有無**に，どれだけ影響するか.』ということを表し，次のような式で書くことが出来る．

$$KL(t_j; t_i) = P(t_j|t_i) \log \frac{P(t_j|t_i)}{P(t_j)}$$
$$+ P(\neg t_j|t_i) \log \frac{P(\neg t_j|t_i)}{P(\neg t_j)}$$

$$W(t_i) = \frac{N_S(t_i)^2}{N_U(t_i)} \times \sum_{t_j \neq t_i} KL(t_j; t_i)$$

3.1.5 χ^2 検定

語の共起のずれを考えると，『単語t_iの**有無**によって，単語t_jの**有無**が，どれだけ影響されるか.』という意味を持つ式で，次の式となる．

$$\chi^2(t_j; t_i) = \frac{\{P(t_j|t_i) - P(t_j)\}^2}{P(t_j)}$$
$$+ \frac{\{P(\neg t_j|t_i) - P(\neg t_j)\}^2}{P(\neg t_j)}$$
$$+ \frac{\{P(t_j|\neg t_i) - P(t_j)\}^2}{P(t_j)}$$
$$+ \frac{\{P(\neg t_j|\neg t_i) - P(\neg t_j)\}^2}{P(\neg t_j)}$$

$$W(t_i) = \frac{N_S(t_i)^2}{N_U(t_i)} \times \sum_{t_j \neq t_i} \chi^2(t_i; t_j)$$

3.1.6 RSV

RSV (Robertson's Selection Value)[5] は，Query Expansion に使われる特徴語選択のための特徴量である．

RSV は以下の式で定義される．

$$RSV_t = w2_t \times (\frac{r_t}{R} - \frac{n_t}{N})$$

ただし，

$$w_t = log(k_4' \times \frac{N}{n_t} + 1),$$

$$w2_t = \alpha \times w_t + (1 - \alpha) \times \log \frac{\frac{r_t+0.5}{R-r_t+0.5}}{\frac{n_t-r_t+0.5}{N-n_t-R+r_t+0.5}},$$

r_t：集合Sの中で，単語tを含む文書数

R：集合Sに含まれる文書数

n_t：データセットの中で，単語tを含む文書数

N：データセットにある全ての文書数

α, k_4'：パラメータ

なお，この特徴量は，単語の共起情報を利用しない．

3.2 実験の概要

3.2.1 実験の手法

NTCIR3 web task[1] は，およそ 1000 万件の Web ページを対象とし，検索課題は 47 個用意されている．各課題において与えられる検索語は 2〜3 個である（図 5 step.1）．

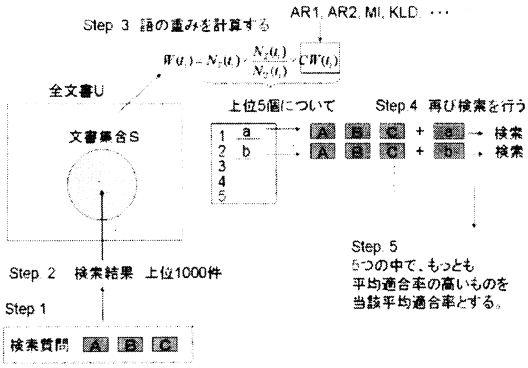

図 5 : 実験の手法.

各検索課題から検索[2]を行い，上位 1000 件を取り出す（図 5 step.2）．この上位 1000 件中において，5 件以上の文書で現れた語だけを残す（文書に現れる頻度が低過ぎないものを残す）．その結果，どの検索課題についても 1 万語前後を得た．また，本実験においては，stop word は除去していない[3]．ここで残った全ての語のペアに関して，共起した文書の数を調べる．そして，共起した文書の数を使って，各式による値を計算する（図 5 step.3）．各手法による結果の上位 5 語 (a, b, c, d, e) を，元々の検索課題 (A+B+C) に 1 つずつ加えて，5 通りの検索 (A+B+C+a, A+B+C+b,・・・, A+B+C+e) を行い（図 5 step.4），最も高かった平均適合率[4]を当該平均適合率とする（図 5 step.5）．最後に，全ての課題についての平均を取り，各式による結果について比較を行う．

3.2.2 実験の意義

予備実験として Query Expansion の手法に則り，上位の複数の語を同時に付け足す実験を行った．その際，同時に付加する語の数を多くする程，平均適合率が悪くなることが分かった．しかし，本手法の目的は，元の検索語が曖昧なために検索結果に含まれてしまう多様なトピックを，良く分離できる語を見つけることである．このような語は，それぞれがひとつのトピックに対応している．したがって，異なるトピックに属し，かつ，互いにそのトピックを他から際立たせるような語を，同時に元の検索語に付け加えて検索すると，検索結果には複数のトピックが混在してしまう．一方，評価用の正解集合は，一つのトピックだけを含むように作られている．それゆえ，本手法の目的を考えれば，評価した結果が悪くなることは容易に推測できる．

そこで，本実験においては，Query Expansion のように抽出された語すべてを一緒に付け足すのではなく，上位にランク付けされた語をそれぞれ別個に元の検索語に追加した．そして，複数得られた検索結果からそれぞれ評価して平均適合率を求め，最も良かった値を当該検索課題の平均適合率とした．なぜなら，最も良い平均適合率を与える語は，評価用の正解集合が含むトピックに一番近いトピックを表わしているはずだからであるこうすれば，少なくとも正解集合が含むトピックを良く分離する語が，今回の手法によって抽出できているかを，実験の結果から判断できる．

もちろん，今回の実験では，正解集合に現れていないトピックについては，それを良く分離する語が取り出せているか，また，対応するトピックをその語がどれだけ適切に表わしているかは，評価できない．しかし，正解集合のトピックについて良い結果が得られると同時に，それとは異なるトピックを表わす語が上位にランク付けされているならば，元の検索語の曖昧さを解消することで現れる別のトピックについても，同様の効果が得られていると期待できる．

3.2.3 実験の結果と考察

表 1 : 得られた平均適合率と平均適合率の上昇率.（元々与えられた 3 語で検索すると平均適合率が 0.1606.）

計算式	平均適合率	上昇率 (%)
UnitWeight	0.1765	9.9
AR1	**0.1847**	15.0
AR2	**0.1899**	18.2
CF	0.1801	12.1
相互情報量 (MI)	0.1829	13.9
KL 情報量 (KL)	0.1733	7.9
χ^2 検定	0.1751	9.0
RSV	**0.1867**	16.3

ベースラインは，元々与えられた 3 語で検索したときの平均適合率であった 0.1606 である．しかし RSV 以外は，共通項である UnitWeight の後ろに各式による重みの和を掛けている．比較対照としての各式の効

[2] OKAPI の式 [2] を使用した．

[3] 検索質問によって stop word となるものは異なってくる．流動的にそのような語を判断できる指標であることが望ましいと考える．

[4] 平均適合率の計算は，TREC の trec_eval を用いた．また，relevance level は rigid である．

表 2：絶滅＋哺乳類＋危機 (baselin は 0.1291)

手法	上位五語	平均適合率
RSV	種	**0.1212**
	生息	0.1105
	生物	0.0762
	野生	0.0923
	動物	0.0724
AR1	レッドデータブック	0.0625
	瀕	0.0739
	危惧	0.1680
	危急	0.1027
	シナノミズラモグラ	**0.2361**
AR2	レッドデータブック	0.0625
	危急	0.1027
	危惧	0.1680
	両生類	0.0747
	ルリカケス	**0.1712**

表 3：スピーカー＋評価＋比較 (baselin は 0.0596)

手法	上位五語	平均適合率
RSV	アンプ	0.0067
	可能	0.0279
	結果	0.0341
	システム	0.0246
	音	**0.0593**
AR1	アンプ	0.0067
	ウーファー	0.0660
	ソフトドームツィーター	0.0154
	スーパーウーファー	0.0177
	バスレフ	**0.0661**
AR2	アンプ	0.0067
	ウーファー	0.0660
	バスレフ	**0.0661**
	サブウーファー	0.0640
	低音	0.0434

果を測るには，UnitWeight の 0.1765 を超えなければ，多少でも効果があったとは言えない．

平均適合率が大きく上昇しているのは，AR1，AR2，RSV の三つであった (表 1)．RSV はもともと，検索要求に適した語群を見つけるために提案されている．しかし，AR1 と AR2 は，検索要求の曖昧さを解消する語群を見つけるために，本研究が提案した．したがって，AR1 や AR2 のスコアが高い単語の中には，評価用の正解集合とは異なるトピックを表す単語も含まれる．なぜなら，検索語が曖昧であるとき，検索結果には複数のトピックが含まれ，正解集合に対応するトピックは，それらのうちの 1 つにすぎないからである．ところが，AR1 や AR2 の場合でも，平均適合率が上昇している．これは，検索結果に含まれる多様なトピックのうち，少なくとも正解集合と同じトピックについては，対応するキーワードを取り出すことができているからである．また，平均適合率が上昇していることより，単に正解集合のトピックに対応する語を取り出すだけでなく，検索の曖昧さを解消するような語を選択できていると言える．この三つの手法（AR1，AR2，RSV）は平均適合率の値では差が無いが，順位付けされた語の一覧を見ると，単語の質が異なっている (表 2〜表 4)．RSV では比較的一般的な意味でも用いられる語が挙がっているが，AR1 や AR2 では検索質問に関連のある特定の分野でのみ使われるような専門的な言葉が多く含まれる．

例えば，「哺乳類，絶滅，危機」という 3 語から (表 2) は，“レッドデータブック[5]” や “ルリカケス[6]”，“シナノミズラモグラ[7]” という固有名詞が，「スピーカー，比較，評価」という 3 語から (表 3) は，“バスレフ[8]” という専門的な用語が，「世界樹＋北欧神話＋名前」という 3 語から (表 4) は，“イグドラシル[9]” という正に北欧神話に出てくる世界樹を指す呼び名が，上位 5 件の中に入っている．

4 関連研究

キーワード抽出（重要語抽出）の手法は，既存の研究でも数多く提案されている．その中でも，語の分布を測り代表性 (representativeness) という指標にあう語を重要とみる手法 [8] や，語の共起に χ^2 検定を利用して重要な語を選ぶ手法 [10] が，語の分布のずれを見るという観点では近い．しかし，全文書における単語の重要度（あるいは各文書中での単語の重要度）で一次元的に並べること考えており，その点では今回の目的と異なる．

今後は，ユーザの入力した語を手がかりに，その曖

[5]環境省が，日本の絶滅のおそれのある野生生物の種についてそれらの生息状況等を取りまとめたもの

[6]1921 年に天然記念物，1993 年に国内希少野生動植物種にそれぞれ指定されている．また環境省のレッドデータブックでは絶滅危惧 II 類（VU）．世界で奄美大島にしか繁殖せず生息しない鳥．

[7]レッドデータブック・哺乳類では，準絶滅危惧（NT）．

[8]正式には「バス・レフレックス型」の略で，密閉型スピーカーボックスに穴（ポート）を開けて，低音域の音を大きく良質にするためのもの．

[9]北欧神話の中で宇宙の中心とされている「世界樹」のことで，トネリコ（モクセイ科の落葉高木）の巨木．

表 4: 世界樹+北欧神話+名前 (baseline は 0.0675)

手法	上位五語	平均適合率
RSV	神	0.0253
	たち	0.0280
	それ	**0.0377**
	歴史	0.0266
	物語	0.0198
AR1	Pandaemonium	0.0586
	イグドラシル	**0.3767**
	ソグネフィヨルド	0.0523
	エッダ	0.0525
	シルマリル	0.0533
AR2	イグドラシル	**0.3767**
	古事記	0.0227
	ギリシア	0.0160
	ノルウェー	0.0203
	フィヨルド	0.0287

昧性を解消するため，あるいは，より詳細な情報を入力するために，システムが情報を提供することが必要となると考えている．そこで，検索語を元に Web 上にあるデータをトピックに分け，曖昧性を指摘できれば，そのいずれであるかをユーザは答えることが可能となるであろう．

トピックに分けるというとクラスタリングを想起するかもしれない．例えば，検索要求に合致するクラスタを選ぶことが検索閲覧の効率化につながるとして作られたシステム:Scatter/Gather[4] がある．しかし，これもまた文書同士の関連性を見るもので，検索要求として入力するのに適した語を探すことは難しい．

可視化という点では，語の共起を可視化してキーワードを抽出する手法：KeyGraph[6][11] や，対話性を重視した検索インターフェースを持ち，特徴語グラフを表示できる DualNAVI[7] といった研究が行われている．しかし一般のユーザに提示する場合には，一見して内容が分かるように表示する必要があると考える．

5 おわりに

本稿では，語の低頻度共起を用いて検索質問の曖昧性を除けるような語を抽出するための手法である分節性（AR1 と AR2) について提案した．特に，AR2 については，情報量的な裏付けを持った定式化であり，また，他手法に比べ格段に良い平均適合率が得られている．今後は，この 2 つの手法による語の選択が，他の手法に対して単語の質が異なるということを実証できるような実験を行う予定である．

また本手法は，検索質問に含まれる曖昧性（多義性）を，それぞれのトピックに分けてユーザに提示できるシステムを最終的な目標としている．トピックに分ける上で有効な手法であることは実験しているが，紙面のため割愛した．今後は，この点に関しても更に研究を進めていく予定である．

参考文献

[1] Oyama K. Ishida E. Kando N. Eguchi, K. and K. Kuriyama. Overview of the web retrieval task at the third ntcir workshop, 2003.

[2] Fang Hui, Tao Tao, and Zhai ChengXiang. A formal study of information retrieval heuristics. In *Proc. of SIGIR 2004*.

[3] Lau, R. Y.K., P.D.Bruza, and D. Song. Belief revision for adaptive information retrieval. In *Proc. of SIGIR'04*, pp. 130–137, 2004.

[4] Hearst Marti and Pedersen Jan. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proc. of SIGIR'96*, pp. 76–84, 1996.

[5] Toyoda Masashi, Kitsuregawa Masaru. Mano Hiroko, Itoh Hideo, and Ogawa Yasushi. University of tokyo/ricoh at ntcir-3 web retrieval task. In *Proc. of the 3rd NTCIR Workshop Meeting*, pp. 31–38, 2002.

[6] Y. Ohsawa, E.B. Nels, , and M. Yachida. Keygraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In *Proc. of IEEE ADL'98*, 1998.

[7] A. Takano, Y. Niwa, S. Nishioka, M. Iwayama, T. Hisamitsu, O. Imaichi, and H. Sakurai. Associative information access using dualnavi. In *Proc. of ICDL'00*, pp. 285–289.

[8] Hisamitsu Toru, Niwa Yoshiki, Nishioka Shingo, Sakurai Hirofumi, Imaichi Osamu, Iwayama Makoto, and Takano Akihiko. Extracting terms by a combination of term frequency and a measure of term representativeness. *International journal of theoretical and applied aissues in specialized communication*, Vol. 6, No. 2, pp. 211–232, 2000.

[9] Yang Yiming and O. Pedersen Jan. A comparative study on feature selection in text categorization. In *Proc. of ICML-97*, pp. 412–420, 1997.

[10] 松尾豊, 石塚満. 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム. 人工知能学会論文誌, Vol. 17, pp. 213–227, 2002.

[11] 大澤幸生, ネルス E. ベンソン, 谷内田正彦. Keygraph: 語の共起グラフの分割・統合によるキーワード抽出. 電子情報通信学会論文誌, Vol. J82-D-I, pp. 391–400, 2 1999.