検索語の曖昧性を解消するキーワードの提示手法

Query Ambiguity Indication Using Infrequent Term Cooccurrences

若木 裕美♥ 正田 備也♦
高須 淳宏♦ 安達 淳♦

Hiromi WAKAKI Tomonari MASADA
Atsuhiro TAKASU Jun ADACHI

既存の検索エンジンではキーワード検索が主流で，数語からなる適切な検索質問を見つけるのが難しい．本稿では，単語の低頻度共起を利用し，検索質問の曖昧性を解消する手法を提案する．この手法は“多種類の単語と共起する単語は，独立したトピックを持ちえない”という仮説に基づき，単語に重みを与え，トピックに分けやすいかどうかの尺度とする．さらに Web データを用いた実験で有効性を確認する．

Conventional search engines are designed mainly for keyword search. Therefore, we have to try many combinations of query terms. This paper presents a query disambiguation method by using infrequent term cooccurrences. Our method weights terms based on the hypothesis that terms appearing with a wide variety of terms cannot establish an independent topic. The weighting is a measure whether the term can articulate a specific topic. Besides we verify the effectiveness of our method by using Web data.

1. はじめに

従来型の検索エンジンでは，ランキングされた結果を見ても所望の情報を得にくい．それは検索エンジンが悪かったのであろうか．そもそも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．現状では，質問者が欲しい情報を得るために，体系立って作られてはいないWeb 上の情報の特性に合わせて「的確に欲しい情報を持ってこられる語」を入力する必要がある．また多くの場合，質問語は1，2語であり，情報量が少ないことも質問処理を困難にしている．その結果、ユーザが必要とはしない内容が検索結果に含まれて提示されていることとなる．異なる内容が含まれていたとしても，検索質問から特定できない場合には，それを排除する術を検索エンジンは持ち合わせていない．多義語に限らず，検索における曖昧性は常に存在すると思われる．

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．これは**検索結果を得る手続きとは別個の手法**であり，検索結果に対して情報の整理、情報の提示を行うことを最終的な目的とするものである．将来的には，集めた語を幾つかの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，そ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る．

2. 分節性（articulateness）

2.1 分節性の仮説

検索質問の曖昧性に対処するため，検索結果に含まれるトピックを分けて提示することを想定する．だが，既存の特徴語抽出でスコアの高い語は，検索結果の文書の多くに現れ，個別的なトピックの手がかりになりにくい．そこで，検索結果に含まれる多様なトピックを識別できる語を抽出する指標として、新たに「分節性(articulateness)」を提案する．分節性の高い語とは，特定のトピックに関する文書に現れやすいが、他のトピックに関する文書には現れにくい語のことだと考える．そして，共起概念を導入してこの考え方を具体化し，次の仮説を立てる：特定の少数の語とよく共起するが，他の多数の語とはあまり共起しない語は，分節性が高い．なお，本稿では，**同じ文書に現れることを共起**と定義する．次に，分節性を定量的に表す二つの定式化 AR1， AR2 を提案する．

2.2 定式化 1：異なり語数に基づく（AR1）

分節性の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．

まず，単語t_iの分節性を，t_iを含む文書に現れる単語の種類の平均数$AvgType(t_i)$を使って定式化することを考える．Sを検索結果として得た文書の集合，$N_S(t_i)$をSの中で単語t_i を含む文書の数，$N_S(t_i \wedge t_j)$をSの中で単語t_i とt_j を含む文書の数とすると，$AvgType(t_i)$ は$\Sigma_{t_j \neq t_i} N_S(t_i \wedge t_j) / N_S(t_i)$ と表すことができる．この$AvgType(t_i)$ が小さいほど単語t_iの分節性が高いとみなす．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，単語t_iがSの中のある程度多くの文書に現れるという条件を加えて，

$$AR1(t_i) = \frac{N_S(t_i)^2}{N_U(t_i)} \cdot \frac{1}{AvgType(t_i)} = \frac{N_S(t_i)^2}{N_U(t_i)} \Big/ \sum_{t_j \neq t_i} \frac{N_S(t_i \wedge t_j)}{N_S(t_i)}$$

という特徴量AR1で単語t_i の分節性を表すことにする．ここで，$N_S(t_i)^2 / N_U(t_i)$という項は，$N_S(t_i) / N_U(t_i)$と$N_S(t_i)$を掛け合わせてつくった項である．$N_S(t_i) / N_U(t_i)$は，検索対象となる文書全部の集合Uの中で単語t_i が現れる文書の数$N_U(t_i)$に対する，Sの中で単語t_i が現れる文書数$N_S(t_i)$の比であり，単語t_i がSにどれだけ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の中で単語t_iが現れる文書数そのもので，単語t_i のSへの関係の度合いを直接表している．

2.3 定式化 2:確率論的観点に基づく（AR2）

もう１つの定式化として、単語t_iの分節性をt_i以外の単語の，集合S での出現頻度と，単語t_i を含む文書集合の中での出現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．分節性の考え方に合うものとして**『その単語が出現することで，他の単語の出現確率が低くなる方向にずれるものほど良い』**というモデルを定式化することを考える．

ここで「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，

♥ 学生会員 東京大学大学院情報理工学系研究科博士課程 hiromi@nii.ac.jp
♦ 正会員 国立情報学研究所 {masada, takasu, adachi}@nii.ac.jp

$$P_S(t_i) = N_S(t_i)/N_S \qquad P_S(t_i \wedge t_j) = N_S(t_i \wedge t_j)/N_S$$

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，$P_S(t_i)$，$P_S(t_i, t_j)$などと書く．

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

とかくとき，$K_j(X)$ は，KL情報量(Kullback Leibler情報量)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 が現れる場合のみのずれであり，$K_j(t_i)$ が大きくなることで表される．$K_j(t_i)$ が大きくなる場合には，次の２つの場合があり得る．

(a) $P(t_j/t_i) \log P(t_j/t_i)P(t_j) < 0$の場合(低くなる方向にずれる)
(b) $P(t_j/t_i) \log P(t_j/t_i)P(t_j) > 0$の場合(増える方向にずれる)

本稿で扱う単語の分節性では，『他の単語の出現確率が低くなる方向にずれる（これをSKLとする）』ものを求めており，前者の項(a) だけが必要となる．そこで分節性は，(a)の条件に合うような単語t_j が多いほど『単語t_i の分節性は高い』と定義でき，KL 情報量の式を変形して定式化することができる．これをAR2 と表すことにする．

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

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

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$cw(t_i, t_j)$ とする．そして，

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

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$CW(t_i)$ と$N_S(t_i)^2 / N_U(t_i)$ 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．

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

特徴語抽出の手法には様々あるが，今回の実験では，AR1，AR2，UnitWeight，CF，相互情報量，KL情報量，$\chi 2$ 検定，RSVの8通りのスコア付け手法を比較した(詳しくは，表4を参照)[7]．

ただし，UnitWeightは，$CW(t_i) = 1$とし，$N_S(t_i)^2 / N_U(t_i)$ 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．またCFでは，AR1と逆の考え方，すなわち『他の語と共起する種類の多い方が良い語である』という指標を定式化し比較対象とした．

また，単語の共起情報を用いないRSV (Robertson's Selection Value) [5]は以下の式で定義される．

$$RSV_{t_i} = w2_{t_i} \times \left(\frac{N_S(t_i)}{N_S} - \frac{N_U(t_i)}{N_U}\right)$$

$$w2_{t_i} = \alpha \times \log\left(k'_4 \times \frac{N_U}{N_U(t_i)} + 1\right) + (1-\alpha) \times \log \frac{\frac{N_S(t_i)+0.5}{N_S - N_S(t_i)+0.5}}{\frac{N_U(t_i) - N_S(t_i)+0.5}{N_U - N_U(t_i) - N_S + N_S(t_i)+0.5}}$$

ただし，α, k_4は，パラメータとする．
RSVは，Query Expansionのための特徴語選択に使われる．

3.2 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1 step.1）．

各検索課題から検索[2]を行い，上位1000 件を取り出す（図1 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．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．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1 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1 step.4），最も高かった平均適合率[3]を当該平均適合率とする（図1 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．

3.3 実験の意義

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．ここで，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，検索結果として複数のトピックのものが混在するだけでなく，検索自体がうまくいかないことも想像される．一方，評価用の正解集合は，１つのトピックだけを含むように作られており，求められるのは1つのトピックに関わるものだけである．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

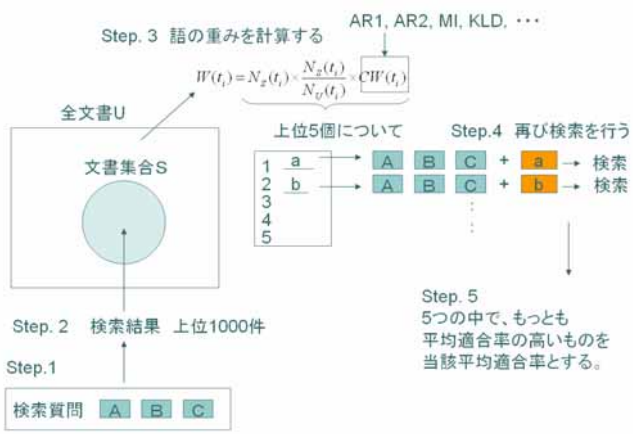

図１ 実験の手法

Fig.1 Experiments procedures.

[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．

[2] OKAPI の式[2] を使用した．

[3] 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 1 得られた平均適合率と平均適合率の上昇率.
Table 1 Overall precisions and their upward rates of the tested term extraction methods.

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
$\chi 2$ 検定	0.1751	9.0
RSV	**0.1867**	16.3

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．もし平均適合率が上昇すれば，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ことを示すといえる．つまり，分離性能については検証できないが，分離性を持つような語であっても検索の性能向上に寄与するかどうかを検証できる実験方法である．

3.4 結果と考察

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ため，比較対象としてUnitWeight の0.1765 を超えなければ，各式で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おり，それと同程度の検索性能上昇に寄与する語がAR1，AR2を用いて得ることができたといえる．

さらに，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3)．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．これは，トピックに分けるための指標となり得ることを示唆している．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[4]”や“ルリカケス[5]”，“シナノミズラモグラ[6]”という固有名詞が，「世界樹+北欧神話+名前」という3 語から(表3) は，“イグドラシル”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[6]や，語の共起に $\chi 2$ 検定を利用して重要な語を選ぶ手法[8] 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(トピックに分ける)目的と異なる．

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．

表 2 検索質問: 絶滅＋哺乳類＋危機 (baselin は 0.1291)
Table 2 Query: “Mammalia”, “extermination”, and “crisis”. (The baseline is 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
	瀬	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 検索質問: 世界樹+北欧神話+名前(baseline は 0.0675)
Table 3 Query: “the World Tree”, “Norse mythology”, and “name”. (The baseline is 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

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

[4] 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

[5] 環境省のレッドデータブックでは絶滅危惧II 類（VU）.

[6] レッドデータブック・哺乳類では，準絶滅危惧（NT）.

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．

[謝辞]

この研究は，文部科学省科学研究費補助金特定領域研究 13224087（2001-2005）の補助を受けています．

[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] F. Hui, T. Tao, and Z. 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] H. Marti and P. Jan. “Reexamining the cluster hypothesis: Scatter/gather on retrieval results”, In *Proc. of SIGIR'96*, pp. 76–84, 1996.

[5] Toyoda M., Kitsuregawa M., Mano H., Itoh H., and Ogawa Y.. “University of tokyo/ricoh at ntcir-3 web retrieval task”, In *Proc. of the 3rd NTCIR Workshop Meeting*, pp. 31–38, 2002.

[6] Hisamitsu T., Niwa Y., Nishioka S., Sakurai H., Imaichi O., Iwayama M., and Takano A.. “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.

[7] Y. Yiming and O. P. Jan. “A comparative study on feature selection in text categorization”, In *Proc. of ICML- 97*, pp. 412–420, 1997.

[8] 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．

若木 裕美 Hiromi WAKAKI

東京大学大学院情報理工学系研究科博士課程在学中．2004 東京大学大学院新領域創成科学研究科修士課程修了．日本データベース学会学生会員．

正田 備也 Tomonari MASADA

2004 年東京大学大学院情報理工学系研究科電子情報学専攻博士課程修了．情報理工学博士．同年，国立情報学研究所プロジェクト研究員．情報検索，データマイニングの研究に従事．情報処理学会，日本データベース学会，各会員．

高須 淳宏 Atsuhiro TAKASU

1989 年東京大学大学院工学系研究科博士課程修了．工学博士．2003 年より同研究所教授．データ工学，電子図書館，機械学習の研究に従事．電子情報通信学会，情報処理学会，人工知能学会, 日本データベース学会, ACM, IEEE 各会員．

安達 淳 Jun ADACHI

1981 年東京大学大学院工学系研究科博士課程修了．工学博士．国立情報学研究所教授．東京大学大学院情報理工学系研究科教授を併任．データベースシステム，データマイニング，情報検索，電子図書館システム等の開発研究に従事．電子情報通信学会，情報処理学会、IEEE，ACM 各会員．

表 4 実験に用いた式の一覧
Table 4 List of formulations for experiments.

手法	$CW(t_i)$	$W(t_i)=\frac{N_S(t_i)^2}{N_U(t_i)}\times CW(t_i)$
AR1	$CW(t_i)=AvgType(t_i)^{-1}=(\sum_{t_j}\frac{N_S(t_i\wedge t_j)}{N_S(t_i)})^{-1}$	$AR1(t_i)=\frac{N_S(t_i)^2}{N_U(t_i)}\times\frac{1}{AvgType(t_i)}$
AR2	$SKL(t_j;t_i)=-P(t_j\mid t_i)\log\frac{P(t_j\mid t_i)}{P(t_j)}+P(\neg t_j\mid t_i)\log\frac{P(\neg t_j\mid t_i)}{P(\neg t_j)}$	$AR2(t_i)=\frac{N_S(t_i)^2}{N_U(t_i)}\times\sum_{t_j\neq t_i}SKL(t_j;t_i)$
UnitWeight	1	$W(t_i)=\frac{N_S(t_i)^2}{N_U(t_i)}\times 1$
CF	$CW(t_i)=AvgType(t_i)=\sum_{t_j\neq t_i}\frac{N_S(t_i\wedge t_j)}{N_S(t_i)}$	$CF(t_i)=\frac{N_S(t_i)^2}{N_U(t_i)}\times AvgType(t_i)$
相互情報量	$MI(t_i,t_j)=P(t_i)\{P(t_j\mid t_i)\log\frac{P(t_j\mid t_i)}{P(t_j)}+P(\neg t_j\mid t_i)\log\frac{P(\neg t_j\mid t_i)}{P(\neg t_j)}\}+P(\neg t_i)\{P(t_j\mid\neg t_i)\log\frac{P(t_j\mid\neg t_i)}{P(t_j)}+P(\neg t_j\mid\neg t_i)\log\frac{P(\neg t_j\mid\neg t_i)}{P(\neg t_j)}\}$	$W(t_i)=\frac{N_S(t_i)^2}{N_U(t_i)}\times\sum_{t_j\neq t_i}MI(t_i;t_j)$
KL 情報量	$KL(t_j;t_i)=P(t_j\mid t_i)\log\frac{P(t_j\mid t_i)}{P(t_j)}+P(\neg t_j\mid t_i)\log\frac{P(\neg t_j\mid t_i)}{P(\neg t_j)}$	$W(t_i)=\frac{N_S(t_i)^2}{N_U(t_i)}\times\sum_{t_j\neq t_j}KL(t_i;t_j)$
$\chi 2$ 検定	$\chi^2(t_j;t_i)=\frac{\{P(t_j\mid t_i)-P(t_j)\}^2}{P(t_j)}+\frac{\{P(\neg t_j\mid t_i)-P(\neg t_j)\}^2}{P(\neg t_j)}+\frac{\{P(t_j\mid\neg t_i)-P(t_j)\}^2}{P(t_j)}+\frac{\{P(\neg t_j\mid\neg t_i)-P(\neg t_j)\}^2}{P(\neg t_j)}$	$W(t_i)=\frac{N_S(t_i)^2}{N_U(t_i)}\times\sum_{t_j\neq t_j}\chi^2(t_i;t_j)$