論 文 Technical Papers

語の活性度に基づくキーワード抽出法

Automatic Indexing Based on Term Activity

松村 真宏 Naohiro Matsumura
東京大学大学院工学系研究科／科学技術振興事業団
Graduate School of Engineering, University of Tokyo / Japan Science and Technology Corporation.
matumura@miv.t.u-tokyo.ac.jp, http://www.miv.t.u-tokyo.ac.jp/~matumura/

大澤 幸生 Yukio Ohsawa
筑波大学大学院ビジネス科学研究科／科学技術振興事業団
Graduate School of Systems Management, University of Tsukuba / Japan Science and Technology Corporation.
osawa@gssm.otsuka.tsukuba.ac.jp, http://www.gssm.otsuka.tsukuba.ac.jp/staff/osawa/

石塚 満 Mitsuru Ishizuka
東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, University of Tokyo.
ishizuka@miv.t.u-tokyo.ac.jp, http://www.miv.t.u-tokyo.ac.jp/~ishizuka/

keywords: PAI, automatic indexing, priming, spreading activation, term activity

Summary

With the increasing number of electronic documents, automatic indexing from a document is an essential approach in information retrieval systems, such as search engines. This paper proposes an automatic indexing method named PAI (Priming Activation Indexing) which extracts keywords expressing assertions of a document. The basic idea is that since an author writes a document for insisting on his/her main point, impressive terms to be born in the mind of the reader could represent the asserted keywords of the document. Our approach employs a spreading activation model to extract keywords based on the activity of terms without using corpus, thesaurus, syntactic analysis, dependency relations between terms, and the other knowledge except for stop-word list. Experimental evaluations are reported by applying PAI to both papers and the archives of a mailing-list.

1. はじめに

文書の電子化が進み，近年ではわざわざ図書館に足を運ばなくても膨大な文書を手軽に閲覧できるようになってきた．しかし，タイトル名や著者名など文書の手がかりが明示的に列挙されていない場合には，キーワードによる検索を行っても所望の文書をなかなか発見できないことはよく経験することであり，せっかくの膨大な文書も有効に活用できているとは言い難い．そこで本論文では，文書から著者独自の主張を表すキーワードを抽出する新しい手法について述べる．似た用語を使う人が多い一分野の中でも，著者独自の主張をキーワードとしておけば，文書の要約やユーザの目前の新しい問題を解決する新しい考えを探すような文書の検索が可能となると考えられるからである．

計算機による自然言語処理はまだ文書の意味を理解する技術まで至っていないので，現在主流となっているキーワード抽出手法は，語の統計情報 (頻度など) や文書の構造上の特徴 (章立てなど) に基づいている．しかし，文書から主張を表すキーワードを抽出することは容易ではない．例えば，繰り返し言及される語は重要な概念を表すと仮定して出現頻度の高い語をキーワードとする方法 [Luhn 57] では，出現頻度の低い著者の主張を表す語を取り出すことはできない．また，同じ分野のコーパスと比較したときに相対的に出現頻度が高い語をキーワードとする方法 [Spark-Jones 72, 長尾 76, Salton 83] だと，文書を特徴づける語が得られるので文書の分類には役立つ [相澤 00] が，そのような語が著者の主張を表すとは一概には言えない．分類によって得られる 1 クラスの文書集合は，既にそのクラスにあたる分野で確立された観点に関連することが多い．これに対し，主張抽出とは，一文書または一著者独自の新しく提案された考えを把握することである．タイトルや見出しなどの位置情報や ‘in conclusion’ などの手がかり表現を利用する手法 [Edmundson 69, 木本 91] でも，一般に文書の表現や構成は著者により大きく異なるので，うまく著者の主張を得られる文書は限定されてしまう．既知の手がかりとなる語の近くから，出現頻度が低くても重要なキーワードを統計的に抽出する手法も提案されているが [Weeber 00]，著者の主張が手がかりとなる語の近くにあるとは限らない．

一方，文書の主張を表す語の抽出にこだわったアルゴリズムに KeyGraph [大澤 99] がある．KeyGraph で

は，文書全体における語の共起関係がその文書の主張に至る筋道を表すと見なして，土台（基礎概念）から導かれる主張を表す語をキーワードとして抽出する．しかし，KeyGraph は文書の構造を土台と主張の 2 層でしか捉えていないので，論旨が幾重にも積み重なって展開されている複雑な論理構造の文書からは，主張を取り出すことは難しい．

では，どのようにすれば著者の主張を表すキーワードを広い範囲の文書から取り出すことができるのだろうか．そもそも主張がない文書から著者の主張を表すキーワードを求めることはできない．そこで本論文では，主張が込められている文書として学術論文と研究の議論を行うメーリングリストの記事を対象として話を進めることにする．すると，著者から読者に主張として伝わる概念を求めることが重要となる．著者の主張を読者がどう受け取るかは，読者の知識，興味などに起因する観点によって異なると考えられるが [奥村 99]，そのような読者の観点を同定することは難しい．そこで本論文では，著者の想定する読者が文書を読むときの記憶の活性状態に着目して，文書を読んだ後に読者の記憶に強い印象を残す語をキーワードとして取り出すことを目指す．

本論文の構成は次のようになる．2 章 で記憶のメカニズムと対応させて著者の主張の定義を行い，3 章 で本研究のベースとなる活性伝搬モデルを紹介する．続く 4 章 で文書から読者の記憶に残る語を抽出するアルゴリズム PAI を提案し，5 章 で学術論文から取り出したキーワードの評価，6 章 でメーリングリストの記事から取り出したキーワードの評価を行う．7 章 で提案手法の関連研究を紹介し，最後に 8 章 で結論を述べる．

2. 著 者 の 主 張

人が文書を読んでその内容を理解できるのは，文書中の語が何らかの形で脳に記憶され，脳がその記憶を解釈するからである．脳がどのようにして記憶を解釈しているのかについて詳細は未知であるが，記憶のメカニズムは徐々に解明されつつある．記憶には，ある語が想起（活性化）されるとその語に関連する語も想起（活性化）されるという**プライミング効果** (Priming Effect) [Lorch 82, Balota 86] があり，また，記憶の想起の早さはその想起頻度に依存することが様々な認知実験によって確認されている [阿部 94].

このプライミング効果は文書の内容を理解する作用にも深く関与していると考えられている．というのも，文書を読み進むにつれて話題が読者の頭の中に展開され，それに伴って記憶が活性化されていく中で文脈を理解し，内容を把握していると考えられるからである．論文の著者もそれに対応して，まず緒論などで研究の背景を述べ，徐々に話題を転換したり膨らませたりしながら読者を著者の展開したい話題へと誘導し，読者の頭の中に話題を理解するために必要な基礎を十分に築いたところで自分の主張を展開する．これにより，著者の主張は読者に理解される．我々は，この考えが自然であると考え，文書の著者はこのように理解を進める読者を想定して執筆するものと仮定している．

そこで我々は，著者のこのような展開に誘導されるようにして読者の記憶に強く残る語，すなわち強く活性化される語を著者の主張を表すキーワードとして取り出す．

3. 活 性 伝 搬

3·1 活性伝搬モデル

2 章 で述べた記憶のメカニズムを近似したものに，**活性伝搬モデル**（Spreading Activation Model）[Quillian 68, Collins 75, Anderson 83] がある．この理論は人の認知的側面から構築されたモデルであり，記憶はノードとして表される語がノードの活性を伝播させるリンクで結びついたネットワーク構造で表される．本研究のベースとなる活性伝搬モデルは式 (1) で表される [Huberman 87].

$$\boldsymbol{A}(t) = \boldsymbol{C} + ((1-\gamma)\boldsymbol{I} + \alpha\boldsymbol{R})\,\boldsymbol{A}(t-1) \qquad (1)$$

ここで，$\boldsymbol{A}(t)$ は活性回数 t の語の活性値を表すベクトル，$\boldsymbol{C}$ はネットワークに注入される活性値を表すベクトル，$\boldsymbol{I}$ は $\boldsymbol{A}(t-1)$ の活性値を $\boldsymbol{A}(t)$ に伝搬させる単位行列，$\boldsymbol{R}$ はネットワークの構造を表す伝搬行列であり，$\boldsymbol{R}$ の i 行 j 列の要素 R_{ij} は語 w_i と語 w_j の関連の強さを表す（対角成分は 0 ）．また，γ は活性値の減衰率を表す減衰パラメータ，α はネットワークが語の活性値に及ぼす影響力の程度を表す伝搬パラメータである．

式 (1) は，外部からの刺激によりネットワーク内のノードに活性が伝搬するモデルである．しかし，文書が読者の記憶に与える効果は，式 (1) の $\boldsymbol{C}$ のように直前までの記憶と無関係に加わるのではなく，直前の記憶から文書の内容によって新たな記憶を導くものであろう．したがって $\boldsymbol{C} = \boldsymbol{0}$ とし，話題の変化に応じてネットワークが構造すなわち語間の関連性まで変化するというモデルをとる．つまり，本論文では伝搬行列を $\boldsymbol{R}(t)$ で表し，活性伝搬モデルを式 (2) で表すこととする．

$$\boldsymbol{A}(t) = ((1-\gamma)\boldsymbol{I} + \alpha\boldsymbol{R}(t))\,\boldsymbol{A}(t-1) \qquad (2)$$

3·2 語のネットワーク

$\boldsymbol{R}(t)$ は，時刻 t における情報の受け手の頭の中での概念間の関係である．文書の読者なら，今，文書のある部分を読んでいることによって得られる語と語の関連性である．本論文の場合，節（節がなければ章）ごとに意味がまとまっており（各まとまりを**セグメント**と呼ぶことにする），それらが順番に読者の頭に入ることによって著者の主張が読者に伝わる．そこで本論文では，各セ

グメントごとに $\boldsymbol{R}(t)$ が決まり，活性伝搬を行う（すなわち式 (1) の t が 1 だけ増える）ことを考える．

各セグメントごとの語のネットワークは，KeyGraph で提案された語の共起グラフを計算するアルゴリズムを利用して求める．詳細は次節にゆずり先に概説すると，まず，各セグメントを理解する上で基本となる概念を表す語を取り出してネットワークのノードとする．次に，ノード間の語の連想の強さを測り，その上位の組にリンクを張る．このようにして，各セグメントを語がリンクで結びついたネットワークで表現し，$\boldsymbol{R}(t)$ を求める．

4. 語の活性度に基づくキーワード抽出

本章では，語の活性度に基づいてキーワードを抽出する新しい手法 **PAI**（Priming Activation Indexing）の具体的な処理について述べる．

4·1 前　処　理

文書は自然言語で書かれているため，そのままでは扱いにくい．そこで前処理として，まず ‘a’ や ‘it’ などの通常はキーワードになり得ない語（ストップワード）を文書から取り除く．ここでは SMART システム [Salton 83] で用いられているストップワードを用いた．次に，語幹が基本的な概念を表し，接尾辞などは統語的な性質を表しているという仮定に基づいて，‘plays’, ‘player’, ‘playing’ などを [Porter 80] の手法で語幹 ‘play’ に縮退する正規化を行う．また，連続する 2 単語の出現頻度があるしきい値（ここでは 3 とした）以上であれば，その単語の組を熟語とみなす [Cohen 95].

なお，ここでは英語で書かれた文書を対象にして説明しているが，日本語で書かれた文書でも形態素解析を行ってわかち書きすれば，以下のアルゴリズムを適用できる．実際に，5 章 では英語論文に，6 章 では日本語のメーリングリストの記事からキーワードを抽出している．

4·2 PAI のアルゴリズム

PAI のアルゴリズムを以下に示す．

Step1) 前処理　4·1 節の前処理を行い，文書からストップワードの除去，接尾辞の処理，熟語の処理を行う．

Step2) 文書の分割　文書をセグメント $S_t(t=1,2,\cdots,n)$ に分割する．

Step3) 伝搬行列 $\boldsymbol{R}(t)$ の導出　各セグメント S_t における語のネットワークの構造を伝搬行列 $\boldsymbol{R}(t)$ として表す．$\boldsymbol{R}(t)$ は次のようにして求める．

まず，各セグメント S_t を理解する上で基本となる概念を表す語として，KeyGraph に倣い S_t における出現頻度の高い語の上位 N_1 個[*1]を選んで $K(t)$ とする．次に，$K(t)$ に含まれる全ての語の組 w_i, w_j $(i \neq j)$ の連想の強さを測るために，S_t 内での w_i，w_j の共起の強さを測る．共起の強さは，式 (3) で表される $co(w_i,w_j)$ で定義する．

$$co(w_i,w_j)=\sum_{s\in S_t}\min(|w_i|_s,|w_j|_s). \quad (3)$$

$|x|_s$ はセグメント s に含まれる文における語 x の出現頻度である．ここで，$K(t)$ の N_1 個の語を冗長なリンクなしに結び合わせるために必要最小限の枝数として，$co(w_i,w_j)$ の上位 N_1-1 個までの語の組 w_i，w_j の間にリンクを張ることにより，語のネットワークを構成する．

$\boldsymbol{R}(t)$ は基本的にはこのネットワークを表す行列であるが，ここで更に，連想関係が強いほどプライミング効果は大きくなることと，語 w_i から 1 本のリンクに伝搬する活性値は w_i に接続しているリンク 1 本 1 本に均等に分かれて伝搬することを仮定する．すなわち，N_1 行 N_1 列の $\boldsymbol{R}(t)$ の i 行 j 列の要素 $R(t)_{ij}$ は $co(w_i,w_j)$ の上位 N_2 までの語の組 w_i，w_j に対して

$$R(t)_{ij}=\frac{co(w_i,w_j)}{w_i\text{に接続しているリンク数}},$$

それ以外の成分については $R(t)_{i,j}=0$ とする．

Step4) 活性伝搬　各セグメント $S_t(t=1,2,\cdots,n)$ について，式 (2) を実行し活性伝搬を行う．ここで $\boldsymbol{A(t)}$ は t 番目のセグメント S_t までに活性化された各語の活性値を表すベクトルであり，伝搬前の各語の活性値の初期値は 1 とする．なお，γ，α の値は適用する文書の種類によって異なるので，詳細は 5 章 と 6 章 に譲ることにする．

Step5) キーワードの抽出　文書の始めから終りまで活性伝搬させて活性値が高くなる語は，2 章 での議論から著者が一貫して強く主張したい語であると仮定する．なお，活性値はそれほど高くならない語でも，重要な概念をつないでいる語は主張を表していることが多い [大澤 99]．重要な概念に溜まった活性値は 1 回の活性化でも近隣に多くの活性値をもたらすと考えられるので，重要な概念をつなぐ語は活性値を活性回数で割った値が高くなる語として得ることができる．そこで，活性値の高い語を**高活性語**，活性値を活性回数で割った値の高い語を**鋭活性語**と定義し，高活性語と鋭活性語を併せて著者の主張を表すキーワードとして取り出す．

5. 論文からのキーワード抽出実験

5·1 各種パラメータの設定

論文における意味の区切り，減衰パラメータ γ，伝搬パラメータ α は次のように与える．まず，論文は節（節がなければ章）単位で意味がまとまっていると考えられ

*1 試行錯誤によって上位 20% の語とした．

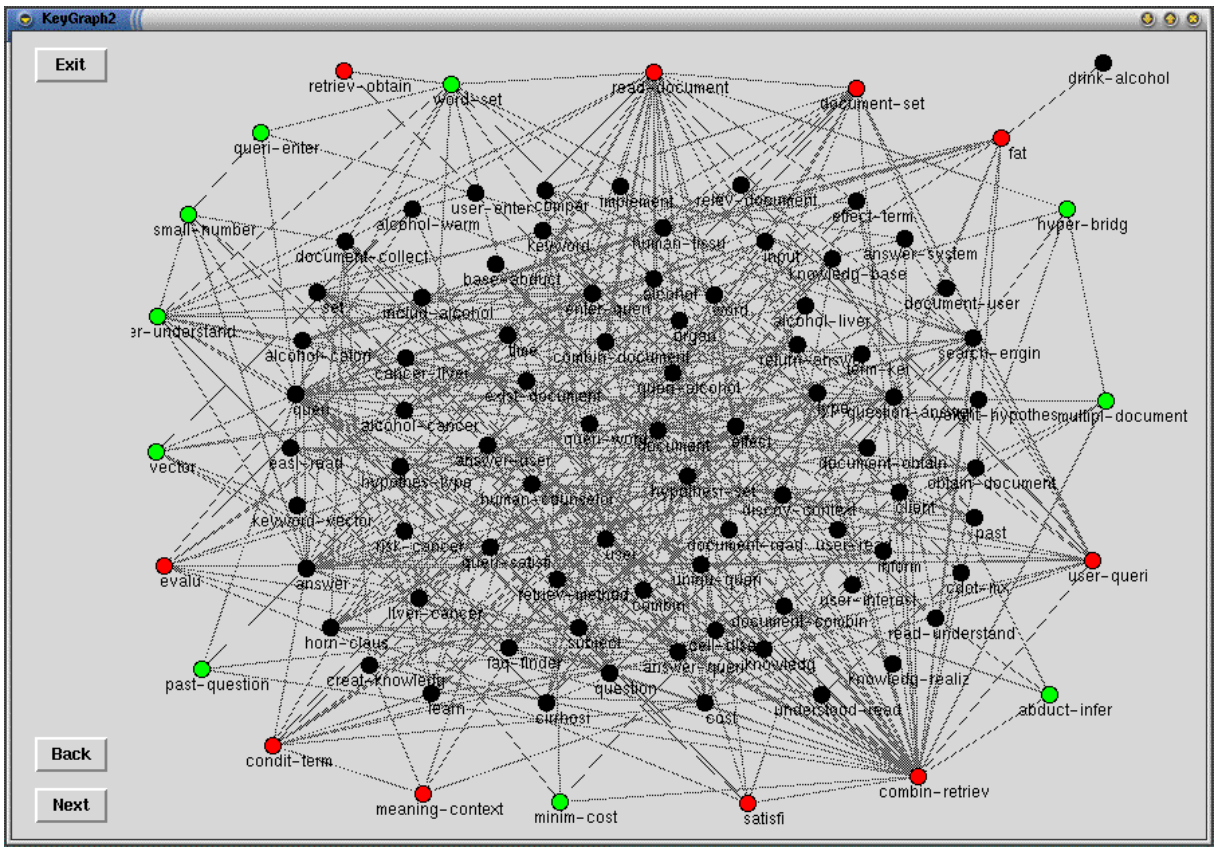

図 1 [Matsumura 00] における語のネットワーク．この図は各セグメントごとに構成される語のネットワークを一つにまとめて表示したものであり，色の薄いノードがキーワードとして得られた語を表している．右端に “multi-document”，右上端に “document-set”，右下端に “combin-retriev”，“abduct-infer”，“past-question”，左上端に “small-number”，左下端に “meaning-context”，“condit-term”，下端に “minim-cost” があるのが見える．

るので，節（節がなければ章）で区切って各セグメントとする．また，一般に記憶は時間が経てば薄れてゆくので，読者にとって語の活性値は文書が長くなれば減衰するかもしれないが [Tanenhaus 79]，著者は読者が読みながら忘れていくことを想定していないと思われる．結局我々が求めたいのは著者の方の主張であるから，一旦増えた活性値が減衰しないように減衰パラメータを $\gamma = 0$ とする．また，$R(t)$ はプライミング効果と語の活性値の伝播量を考慮しているが，この際にさまざまな仮定に基づいているので，最終的には伝搬パラメータの値を予備実験により求めて $R(t)$ を調整する．ここでは，本実験で用いていない 5 つの論文を用いて行った予備実験で最も結果の良かった $\alpha = 1$ を伝搬パラメータとする．

5·2 論文からのキーワード抽出例

論文の例として，[Matsumura 00] を PAI にかけて取り出した高活性語を 表 1 に，鋭活性語を 表 2 に示す．PAI の語のネットワークは 図 1 のようになる．また，比較のために KeyGraph による出力を 表 3，TFIDF [Salton 83] による出力を 表 4，TF [Luhn 57] による出力を 表 5 に示す．なお [Matsumura 00] は，一文書では満たせないユーザの興味を満たすために複数の文書を組み合わせて取り出すことを提案している論文であり，単語数は 5341，単語の種類は 1230 種であった．

著者によると，この論文で最も言いたいことは “combination retrieval” と “document set”（“multiple documents” も同様の意味で用いられている）であり，頻度順位は 1 位と 60 位（“multiple documents” は 188 位）である．“combination retrieval” は頻度順位が 1 位ということもあっていずれの手法でも上位に出現しているが（KeyGraph では 13 位に表れている），“document set”（“multiple documents”）は KeyGraph，TFIDF，TF のいずれでも取り出せていない．また，それ以外の語を見ても，高活性語には “meaning context”（155 位），“conditional term”（70 位），鋭活性語には “abductive inference”（176 位），“small number”（143 位），“minimal cost”（175 位），“past question”（208 位）など出現頻度は低いけれども著者にとって特に重要な語も取り出されていることが分かる．

TFIDF では他の多くの論文にも表れるような語の重要度は低くなるので，そのような語は実際は重要であってもキーワードとして拾うことは難しい．したがって，上記で挙げた “abductive inference” など，[Matsumura 00] のキーワードとしてふさわしいのにも関わらず，同じ人工知能のうち，非単調推論とその応用を扱う分野の論文にも多く表れているので語の重要度は低く見なされ

表 1 PAI による高活性語

順位	語	活性値	頻度順位
1	user queri	1125.0	39
2	read document	345.65	9
3	fat	305.11	11
4	satisfi	207.21	28
5	evalu	161.14	33
6	retriev obtain	130.11	522
7	document set	93.123	60
8	meaning context	89.451	155
9	condit term	85.107	70
10	combin retriev	68.111	1

表 2 PAI による鋭活性語

順位	語	活性値 / 活性回数	頻度順位
1	abduct infer	4.9564	176
2	small number	4.8493	143
3	user understand	4.4820	44
4	minim cost	4.1231	175
5	multipl document	3.7528	188
6	queri enter	3.7247	177
7	vector	3.1492	94
8	word set	2.3746	51
9	hyper bridg	2.3716	171
10	past question	2.2993	208

表 3 KeyGraph によるキーワード

順位	語
1	document
2	alcohol
3	user
4	query
5	doc
6	weights
7	subject
8	fat
9	understandable
10	types

表 4 TFIDF によるキーワード

順位	語
1	combin retriev
2	document
3	queri
4	user
5	answer
6	read document
7	alcohol
8	keyword
9	question answer
10	answer queri

表 5 TF によるキーワード

順位	語
1	combin retriev
2	document
3	user
4	queri
5	answer
6	knowledge
7	obtain
8	word
9	read document
10	alcohol

てしまう．また，“abductive inference” は出現頻度が低いので，単純に出現頻度の高い語を重要であると見なす TF でもキーワードに選ばれることは難しい．一方，“abductive inference” は，最適な “document set” を “combination retrieval” によって取り出す手法として用いられており，実際に論文中においてもそのように記述されている．つまり，“abductive inference” は非常に重要な語と共起している語であり，したがって語のネットワーク構造を利用する PAI が鋭活性語として “abductive inference” の重要性を測れたのだと考えられる．

さらに，KeyGraph では文書全体における頻出語を用いて語の共起グラフを形成し，文書全体の構造を大局的に捉えようとするのに対し，本手法ではセグメントごとに語のネットワークを作成するので，文書の構造をセグメント単位で捉えることができる．“abductive inference” は部分的にしか現れない語であったために，KeyGraph の大局的な視点では捉え切れなかったのだと言える．

5·3 評 価 実 験

PAI の評価を行うために，KeyGraph，TFIDF，TF との比較を行った*2．後述の被験者は全部で 6 名 (博士課程の学生 3 名，研究者 3 名) であり，対象となった論文は被験者が原著として英語で書いた 23 篇である．

実験は以下の手続きにより行った．まず，各論文から PAI，KeyGraph，TFIDF，TF によりキーワードを 15 語ずつ抽出した．PAI では高活性語の上位 10 語，鋭活性語の上位 5 語を併せてキーワードとした．また，TFIDF に用いたコーパスは Journal of Artificial Intellifence Research *3 から入手した 93 年（Volume 1）から 2001 年（Volume 14）までの論文計 166 篇から作成した．次に，各論文について抽出したキーワードを論文の原著になっている被験者に見せ，文書の主張としてふさわしい語をチェックしてもらった．precision（取り出したキーワードのうち，主張を表していると判断されたキーワードの数の割合）と，主張と表していると判断されたキーワードの平均出現頻度で評価でした結果を 表 6 に示す．なお，当初は recall（主張を表していると判断されたキーワードのうち，取り出せたキーワードの割合）も評価するつもりで著者に主張を表す語を全て書き出してもらうことを試みたのだが，分野を表すようなキーワードしか得られず，主張を表すキーワードは 2，3 語程度しか取り出すことができなかった*4．しかし，著者が主張に挙げ

*2 同じ条件で比較するために，4·1 節 の前処理を行った文書に対してキーワードを求めた．

*3 http://www.cs.washington.edu/research/jair/home.html

*4 通常，論文の著者は関連分野を示すためのキーワードを書くように学会などから要請されており，「主張を表す語を示せ」と

表 6 実験結果

	PAI	KeyGraph	TFIDF	TF
Precision	0.56	0.45	0.63	0.55
平均出現頻度	14.3	17.9	19.4	24.1

ていないキーワードでも著者に聞いてみれば主張を表すキーワードであった例が多数あった．このような理由から，recall を評価することは難しいと判断し，precision と平均出現頻度による評価を行った．

一般に単語の出現頻度 f と出現順位 r の間には $r \cdot f = C$（C は定数）が経験的に成り立つことが知られており（Zipf の法則 [Zipf 49]），また経験的には出現頻度の高い語がキーワードである割合も高い．したがって，出現頻度の低い語の中から重要な語を見つけることは，出現頻度の高い語の中から重要な語を見つけることより難しい問題となる．PAI は precision を見れば TFIDF よりも低く TF と同程度であるが，取り出せたキーワードの平均出現頻度は TFIDF，TF よりもかなり低くなっており，この精度をコーパスを使わずに達成していることを考えると，PAI は文書の話題の流れを捉えて著者の主張を取り出せていると評価することができよう．

しかし，TF の結果から分かるように，論文では頻度の高い語がキーワードである確率が高い上に，古くからある分野の論文ではコーパスを用意しやすいので，PAI や KeyGraph ならではの特徴が表れにくい実験となっている．そこで次章では，あるテーマで議論を行うメーリングリストの記事からキーワードを抽出する実験を行う．今回用いたメーリングリストの記事は，データマイニング，リスクマネージメント，社会心理学，マーケティングなど様々な分野をまたぐチャンス発見が話題になっているためにコーパスを用意できない．したがって，従来の手法では適切にキーワードを取り出すことは難しい問題となっている．

6. メーリングリストの記事からのキーワード抽出実験

6·1 メーリングリストの記事のテーマとその特徴

ここでいうメーリングリストは，同じ分野，テーマに興味を持った同士が互いにメールをやり取りする場であり，そこでは様々な意見が交わされ，新しいトピックが湧き出てくる．ここでは，著者らが属する科学技術振興事業団の「自然現象・社会動向からの予兆発見とその利用」チームが作成したチャンス発見に関するメーリングリストを用いる．実験に用いたメールは，このメーリングリストが発足した日の 2000 年 11 月 14 日から実験を行った 2001 年 7 月 5 日までにやり取りした全てのメールの本文 215 通のうち，最初の 5 通のテストメールを除いた 210 通であり，記事全部で 10241 行，353497 バイトであった．なお，前処理として茶筌 [茶筌] により形態素解析を行い，名詞，形容詞，副詞，動詞，未知語だけを残し，メーリングリストの登録者の名前とシグネチャに含まれる単語/記号をストップワードとして除いた．

いわれることに不慣れなことが 1 つの原因とみている．

表 7 PAI による高活性語

順位	語	活性値	頻度順位
1	チャンス発見	93184	8
2	提案	93076	523
3	feature selection	41935	1101
4	知識発見	41935	653
5	結果	16773	5
6	グループ	16653	851
7	tel	16554	457
8	JST	12828	44
9	可能性	12681	381
10	仮説	12085	302

表 8 PAI による鋭活性語

順位	語	活性値 / 活性回数	頻度順位
1	論文	3989.6	2151
2	データマイニング	1712.8	432
3	月	575.47	42
4	収集	543.95	241
5	思う	467.36	496
6	わかる	442.86	652
7	意思決定	378.63	1429
8	場所	364.63	292
9	主体	313.47	444
10	開発	294.24	478

6·2 各種パラメータの設定

メーリングリストの記事の集合から意味のまとまりとなるセグメントを正確に見つけることは難しい．例えば，メールのヘッダを見れば新規メールかどのメールへの返信メールかは分かるが，返信メールでも話題が次から次へと移ってゆく場合も多く，また実際には返信ボタンを押して内容が全く異なるメールを作成することも多いので，スレッドを解析しても意味のない構造が出ることが多い．そこで本論文では，メールが送られた時間に沿って便宜的に 10 通ごとに 1 つのセグメントと見なした．

なお，ここで用いたメーリングリストの記事は約 8ヵ月の間にやり取りされたものなので，時間の経過に伴って過去の記事の記憶は薄れていると考えられる．また，メーリングリストでの議論は，論文の場合と比べると発散したものになる．ここでは様々な α，γ の組み合わせの中から最も結果の良かった伝搬パラメータ $\alpha = 0.9$，減衰パラメータ $\gamma = 0.1$ による結果を示す．α と γ の値が 5 章 の論文からのキーワード抽出実験の場合と異なってい

表 9 TF によるキーワード

順位	語
1	思う
2	人
3	データ
4	結果
5	下さる
6	先生
7	チャンス発見
8	まつむら
9	ランク
10	考える
11	co
12	そう
13	やる
14	見る
15	下記
16	メール
17	ページ
18	コミュニティ
19	ところ
20	WWW

のは，このような論文とメーリングリストの特徴の違いを反映しているからだと考えている．

メーリングリストの全記事から PAI によって取り出された高活性語を 表 7，鋭活性語を 表 8，TF による結果を 表 9 に示す．なお，このメーリングリストにふさわしいコーパスを用意することができないので TFIDF ではキーワードを求めることができず，また KeyGraph は日本語を扱えないので*5，ここでは TF の結果と比較することにする．

6·3 キーワード抽出結果の考察

ここでは，PAI，TF が出力したキーワードの評価を行う．まず，表 7 の PAI で高活性語として得られたキーワードを見ると,「チャンス発見」「feature selection」「知識発見」「結果」「グループ」「JST」「可能性」「仮説」が以下にしめすメーリングリストの要旨 1 から主張を表すキーワードとしてふさわしい．

> **[要旨 1]** 「チャンス発見」において扱う対象は稀な事象まで含むので，これまでの KDD (データベースからの「知識発見」) ではノイズと見なされていたようなデータでも，チャンスとなる「可能性」がある限り捨てることができない．したがって，いかにしてデータから稀でも重要な事象を拾うかという「feature selection」が非常に大事となる．それらのデータを解析した「結果」を「グループ」ディスカッションにより解釈してチャンスになるかもしれない「仮説」を導く手法の確立が,「JST」での我々のチームに課せられた課題である．

また，表 8 の鋭活性語として得られたキーワードのうち,「データマイニング」「意志決定」「主体」も以下の要旨 2 から主張を表すキーワードとしてふさわしい．

> **[要旨 2]** チャンスは，主観的な仮説を評価することによりその価値を評価する．従来の「データマイニング」とチャンス発見の違いはここにあり，チャンスは主観の中から見出されるものであり,「意志決定」する「主体」(人) が考えて仮説を解釈することがチャンス発見の根幹をなす．

表 9 の TF によるキーワードは，上記の要旨 1,2 で出てきた「人」「データ」「結果」「チャンス発見」「考える」の他にも,「コミュニティ」は以下の要旨 3 から主張を表すキーワードとしてふさわしい．

> **[要旨 3]** i-mode が携帯電話を使う人たちとインターネットを使う人たちのニーズを満たし，ハイブリッド自動車が環境問題を考える人たちと自動車を使う人たちのニーズを満たしたように，チャンスは複数の「コミュニティ」が出会う場に生まれる．

要旨 3 はチャンス発見に関するミーティングを始めた当初から議論されつづけている話題であり，最近では要旨 1，要旨 2 で述べたように，より具体的な話に発展してきている．したがって，主張を表すキーワードとしては最近の話題を反映した要旨 1，要旨 2 に表れるキーワードの方が相応しい．しかも，PAI により得られたキーワードの頻度順位は 「feature selection」は 1101 位,「知識発見」は 653 位,「グループ」は 737 位,「可能性」は 381 位,「仮説」は 302 位,「データマイニング」は 432 位,「意志決定」は 1429 位,「主体」は 444 位であり，頻度が低いので従来の手法では抽出し難いキーワードである．

PAI，TF ともにストップワードリスト以外の外部知識は一切用いておらず*6，語の統計情報しか用いていない．したがって，キーワードになり得ないような語がノイズとして混ざってしまう問題はどうしても避けられないが，表 7，表 8，表 9 を比べれば明らかなように，PAI では TF に比べてノイズが少ない．実際に，PAI では高活性語，鋭活性語合わせて 20 語のうち 11 語が主張を表すキーワードとして相応しいが，TF では 20 語のう

*5 6·1 節と同様の前処理を行えば日本語の文書に KeyGraph を適用することも不可能ではないが，KeyGraph は英語の論文から主張を表すキーワードを取り出すようにチューニングされているので，ここでは用いないことにした．

*6 日本語の文書を形態素に分けるときには辞書を用いているが，これらの辞書は PAI，TF のアルゴリズムとは関係がないので，こう表現した．

ち 6 語 だけに留まる．

以上の考察より，PAI は文脈を捉えることによって，質も精度も良く主張を表すキーワードを取り出すことができると結論づけることができる．

7. 関 連 研 究

以前から文書の内容を把握するための手がかりとして文脈や活性伝搬が注目されている [Waltz 85, Grosz 86]．例えば，Norvig [Norvig 89] は文書理解における人の推論過程を意味ネットワーク上の活性伝搬により説明している．また，小嶋 [小嶋 91] らは知識と文脈を利用して文書の意味解釈の曖昧性を解消するモデルを提案している．Mani [Mani 97] らは意味ネットワーク上の活性伝搬により複数の文書の要約を試みている．

これらの手法はいずれも，人が文書を読む時の記憶の活性状態に着目しているという点は我々と同じである．しかし，我々は著者が読者に考えを伝えようとする意図が文書の流れに表れていることに着目しており，文脈すなわち語の共起構造により活性化される語を著者の主張を表す語として取り出している点で従来手法と異なる．

8. ま と め

文書は著者が自分の考えを読者に伝えるために書かれるものであるから，文書を読んだ後に強く読者の記憶に印象を残すような語が著者の主張を表すキーワードとしてふさわしいであろう．このような考えに基づき，本論文では，著者の想定する読者が文書を読むときの記憶の活性状態に着目して，文書を読んだ後に読者の印象に強く残る語をキーワードとして取り出す PAI を提案した．そして，論文からのキーワード抽出，研究の議論を行うメーリングリストの記事からのキーワード抽出実験を行い，PAI が出現頻度が低くても重要な語を的確に取り出せることを示した．我々はそのような語は人の意志決定にとって重要な役割を担うと考えており，その効果を確認するために今後は会議録や談話録などの文書に PAI を適用することを考えている．

◇ 参 考 文 献 ◇

[阿部 94] 阿部純一, 桃内佳雄, 金子康朗, 李光五: 人間の言語情報処理, サイエンス社, 1994.

[相澤 00] 相澤彰子：語と文書の共起に基づく特徴量の数量的表現について，情報処理学会論文誌 Vol. 41, No. 12, pp. 3332–3343, 2000.

[Anderson 83] J.R. Anderson: A spreading activation theory of memory, *Journal of Verbal Learning and Verbal Behavior*, 22, pp. 261–295, 1983.

[Balota 86] D.A. Balota and R.F. Lorch: Depth of automatic spreading activation: Mediated Priming Effects in Pronunciation but not in Lexical Decision, *Journal of Experimental Psychology: Learning, Memory, Cognition*, 12, pp. 336–345, 1986.

[茶筅] 茶筅, `http://chasen.aist-nara.ac.jp/`

[Cohen 95] J. Cohen: Highlights: Language- and Domain-Independent Automatic Indexing Terms for Abstracting, *Journal of American Society for Information Science*, 46, pp. 162–174, 1995.

[Collins 75] A.M. Collins and E.F. Loftus: A Spreading-Activation Theory of Semantic Processing. *Psychological Review*, 82, pp. 407–428, 1975.

[Edmundson 69] H. Edmundson: New Methods in Automatic Abstracting, *Journal of ACM*, 16(2), pp. 264–285, 1969.

[Grosz 86] B.J. Grosz and C.L. Sidner: Attention, Intentions, and the Structure of Discourse, *Computational Linguistics*, Vol. 12, pp. 175–204, 1986.

[Huberman 87] B.A. Huberman, T. Hogg: Phase transitions in artificial intelligence systems, *Artificial Intelligence*, 33, pp. 155–171, 1987.

[木本 91] 木本晴夫: 日本語新聞記事からのキーワード自動抽出と重要度評価, 電子情報通信学会, Vol. 74-D-I, No. 8, pp. 556–566, 1991.

[小嶋 91] 小嶋秀樹, 古郡廷治: テキスト解釈の曖昧性を知識と文脈によって解消する計算モデル, 情報処理学会論文誌 Vol.32, pp. 1366–1373, 1991.

[Lorch 82] R.F. Lorch: Priming and searching processes in semantic memory: A test of three models of spreading activation, *Journal of Verbal Learning and Verbal Behavior*, 21, pp. 468–492, 1982.

[Luhn 57] H.P. Luhn: A Statistical Approach to the Mechanized Encoding and Searching of Literary Information, *IBM Journal of Research and Development*, Vol. 1, No. 4, pp. 309–317, 1957.

[Mani 97] I. Mani and E. Bloedorn: Multi-document Summarization by Graph Search and Matching, *Proceedings of the 14th National Conference on Artificial Intelligence*, pp. 622–628, 1997.

[Matsumura 00] N. Matsumura and Y. Ohsawa: Combination Retrieval for Creating Knowledge from Sparse Document Collection. *Proceedings of Discovery Science*, pp. 320–324, 2000.

[長尾 76] 長尾真, 水谷幹男, 池田: 日本語文献における重要語の自動抽出, 情報処理, Vol. 17, No. 2, pp. 110–117, 1976.

[Norvig 89] P. Norvig: Marker Passing as a Week Method for Text Inferencing, *Cognitive Science*, Vol. 13, No. 4, pp. 569–620, 1989.

[大澤 99] 大澤幸生, Nels E. Benson, 谷内田正彦: KeyGraph: 単語共起グラブの分割・統合によるキーワード抽出, 電子情報通信学会論文誌 J82-D1, No.2, pp. 391–400, 1999.

[奥村 99] 奥村学，難波英嗣：テキスト自動要約に関する研究動向，自然言語処理，Vol.6，No.6，pp.1–26，1999.

[Porter 80] M.F. Porter: An Algorithm for Suffix Stripping, *Automated Library and Informations Systems*, Vol. 14, No. 3, pp. 130–137, 1980.

[Quillian 68] M.R. Quillian: Semantic Memory, *Semantic information processing, MIT Press*, pp. 227–270, 1968.

[Salton 83] G. Salton and M.J. McGill: Introduction to Modern Information Retrieval, *McGraw-Hill*, 1983.

[Spark-Jones 72] K. Spark-Jones: A Statistical Interpretation of Term Specificity and Its Application in Retrieval, *Journal of Documentation*, Vol. 28, No. 5, pp. 111–121, 1972.

[Tanenhaus 79] M.K. Tanenhaus, J.M. Leiman and M.S. Seidenberg: Evidence for Multiple Stages in the Processing of Ambiguous Words in Syntactiv Contexts, *Journal of Verbal Learning and Verbal Behavior*, 18, pp. 427–440, 1979.

[Waltz 85] D.L. Waltz and J.B. Pollack: Massively parallel parsing: A Strongly Interactive Model of Natural Language Interpretation, *Cognitive Science*, 9, pp. 51–74, 1985.

[Weeber 00] M. Weeber, R. Vos and R. H. Baayen: Extracting the Lowest-Frequency Words: Pitfalls and Possibilities, *Computational Linguistics*, Vol. 26, No. 3, pp. 301–317, 2000.

[Zipf 49] G. K. Zipf: Human Behavior and the Principle of Least Effort, *Addison-Wesley*, 1949.

〔担当委員：外山勝彦〕

2001 年 8 月 20 日 受理

謝　辞

著 者 紹 介

松村 真宏（学生会員）

1998 年 大阪大学基礎工学部システム工学科卒業．2000 年 同大学院修士課程修了．現在，東京大学大学院工学系研究科博士課程在学中．2000 年より科学技術振興事業団リサーチスタッフ．最近は人間の意思決定のプロセスに興味がある．2001 年 人工知能学会 MYCOM 優秀プレゼンテーション賞受賞．

大澤 幸生（正会員）

1990 年 東京大学工学部卒業．1995 年同大学院博士課程修了．博士 (工学)．大阪大学基礎工学部助手を経て 1999 年より筑波大学社会工学系助教授，現在に至る．2000 年より科学技術振興事業団研究者を兼任，予兆発見研究に従事．情報処理学会，AAAI，IEEE などの会員．人工知能学会では 1994 年，1999 年全国大会優秀論文賞，1998 年論文賞受賞．

石塚 満（正会員）

1971 年東京大学工学部電子卒業．1976 年同大学院博士課程修了．工学博士．同年 NTT 入社，横須賀研究所．1978 年東京大学生産技術研究所助教授．1992 年工学部電子情報工学科教授．2001 年より情報理工学系研究科電子情報専攻．研究分野は人工知能，知識処理，マルチモーダル擬人化エージェント，ネットワーク化知的情報環境．IEEE，AAAI，情報処理学会，人工知能学会，映像情報メディア学会，画像電子学会等の会員．