文献テキストからのキーワードマイニングと内容検索への応用

土橋 喜
愛知大学現代中国学部

Keyword Mining from Text for Documents Browsing

Konomu Dobashi
Aichi University

目　　次

1. はじめに

最近のインターネットには有益なさまざまな情報が公開されている．中には研究成果を発表した論文が電子ジャーナルの形式で公開されているものなども数多くある．これら公開された文献はダウンロードが許されている場合が多く，著作権の範囲内であれば個人的に収集して再活用することが可能である．ダウンロードもWebブラウザを使えば手軽に行えるようになっており，このような電子的な文献の提供方法は今後も増加すると予想される．

ところで収集した文献が比較的少ないあいだは，自分の記憶にたよりながら一つ一つの文献をブラウザなどに表示して手軽に読むことができる．ダウンロードしたときに自分でわかりやすいファイル名を付けておけば，混乱することなく目的の文献を見出すこともできる．

しかし重要なキーワードがどの文献のどのあたりに現れていたかということになると，人間の記憶の限界を感じる場合が出てくる．

また収集した文献の数が多くなるといろいろ不都合が出てくる．例えば数十から数百の文献を集めた場合，文献のタイトルやそれを格納したファイル名だけでも正確に覚えていることは困難になってくる．ダウンロードした時に自分でファイル名を付け替えたりする場合も多く，ファイル名が必ずしも内容を正確に反映していなかったり，重要なキーワードを覚えてはいるが，そのキーワードがどの文献のどのあたりにあったかなどは思い出せなくなる場合も増える．

このような場合に，収集した文献がタイトルや著者名などから検索できたり，重要な用語（キーワード）や専門用語などがマウスのクリックだけ一覧できたり，それらの用語から元の文献が検索できるようなシステムがなければ，収集した文献の整理や内容の検索などは不可能になってくる．

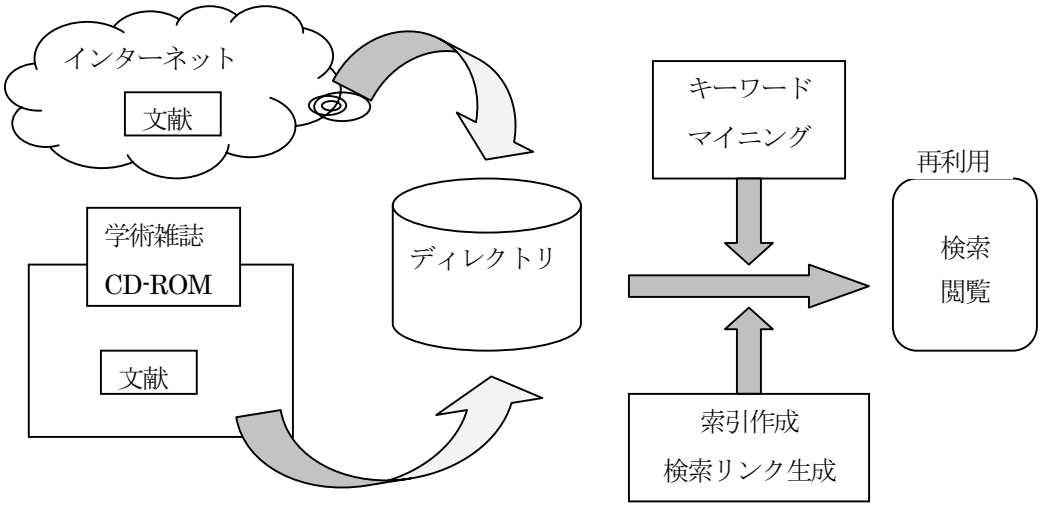

図 1　システムの位置づけ

インターネットにおける情報の氾濫は，今後このような文献整理や内容検索の機能を持つシステムが，個人レベルでも重要になっていることを示している．そこでここでは主にインターネットから収集した文献を対象に，個人的な再利用を前提として，手軽に使える文献整理と内容検索の機能を提案する（図 1）．なおここで提案する方法は，著作権法の範囲内で公開された情報の再利用を個人的に行うものであり，収集した情報を再度インターネットに公開することではない．

2. キーワードの重要性

インターネットに公開される文献には，テキスト，音声，静止画，動画などいくつかの種類があるが，ここではディジタル形式で収集した文献の整理支援を目的にすることから，テキストまたは HTML でタグ付けされた文献を対象として考えていく．

最近のインターネットにおける goo や Google などの検索エンジンでは，ページ内で使われている単語から検索する全文検索機能が使われている．これらの検索エンジンではいろいろな工夫を行い Web ページに付けられたタイトルを取り出し，リンクを作成して一覧表示している．しかしタイトルについては，Web ページの作成者が常に適切なタイトルを付けるとは限らないため，抽出されたタイトルが意味不明のこともたびたびある．タイトルに出ているだろうと思われる単語を入力したにもかかわらず，Web ページの内容とは全く無関係のタイトルが出てくる場合がこれであり，Web ページの作成者が<title></title>タグに適切なタイトル名を書かなかったためである．

例えば Google を使って「情報処理」と検索語を入力したときに，「INTAP HOME PAGE」というタイトルのページが検索されるような場合である．検索エンジンが全文検索を行っている場合は，検索のために入力した文字列がタイトルの部分に出現していなくても，タグ以外の文献のどこかに使われていればヒットする場合が多い．上の「INTAP HOME PAGE」の例がその例であり，ページを開いて見ると「情報処理」という用語が使われていることがわかる．ヒットした文献のタイトルには，必ずしも入力した単語が含まれていない場合もあるため，Google などではその単語が出現する文章をタイトルと一緒に表示するなどの工夫が行われている．このように Web ページの検索ではタイトルだけでは必ずしも有効でない場合があり，文献中のキーワード（重要語）を効果的に検索する意義は益々高まっている．

またキーワードを適切に抽出することができれば，それらを使って文献内を検索するリンクを生成し，検索機能を実現することができるようになる．またこの研究では取り上げないが，情報検索の分野では文献中におけるキーワードの出現頻度から，文献の類似度を判定し検索に活用することが行われている．

どれがキーワードかを選定する場合，人間が行うときは手作業で内容を全部読む必要があるが，文献の数が多くなると不可能である．そのため従来はタイトルや著者名など文献の重要かつ限られた部分からキーワードを抽出していたが，近年のコンピュータとソフトウエアの性能向上により，人手に頼らず文献全体からシステムで行うことが可能になり，関連した研究も多い[Nagao 76，Ohsawa 99]．

このような流れの中で，本稿では重要なキーワードをシステムで自動的に取り出すキーワードマイニン

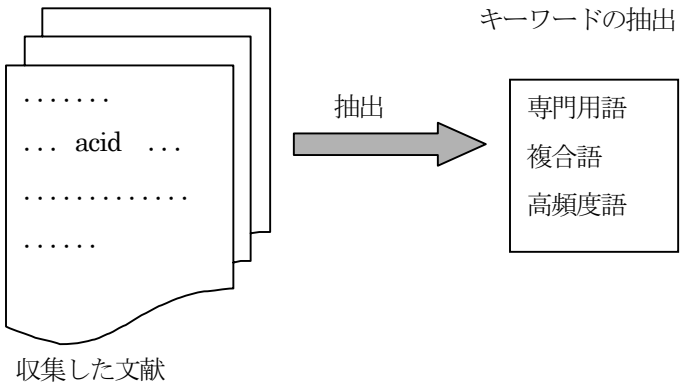

図 2　収集した文献からのキーワード抽出

グの仕組みを開発し，抽出されたキーワードを使って，文献内容を検索するリンクの生成を目標としている．

3.　抽出するキーワードと処理対象

文献のどの部分が重要かということは，人間が読んでもなかなか判断が難しいところがあり，文献が扱っている内容にも関わる問題である．さらに読み手の主観的な見方や何のために読むかという必要性も加わって来るため，どこが重要な部分であるかという判断に，人によってばらつきが出てくることが当然のように予想される．

この問題を突き詰めていくと，文献の意味内容を理解する必要性が指摘され，コンピュータに意味を考えさせるような極めて困難な課題となり，コンピュータでは扱いにくい問題になりかねない．ここでは文献のなかでも比較的重要な部分が特定しやすいものから取り組んで行く．そこで専門用語，出現頻度の高い用語すなわち高頻度語および複合語の 3 つをキーワードとして扱うことにする（図 2）．

従って扱う文献も文学作品などではなく，科学技術関係の文献に近いものを扱い，ここでは地球環境問題の文献を対象に考える．地球環境問題の文献を扱う理由としては，対策が緊急を要しているにも関わらず，問題が大規模で複雑なため多くの分野に関連しており，問題点を整理して解決策を策定するための支援システムが必要なためである．

また言語としてはコンピュータでの扱いやすさを考え，英語を対象として議論を進める．自然言語を扱う場合には，文字だけではなく文法なども考慮する必要があり，言語が異なるとそれに対応する処理を行わねばならないが，英語の場合はこれまでの研究成果の活用が期待できるのも理由である[JEITA 00, Rajman 97].

さらに最近のインターネットでは，さまざまな形式のファイルが公開されているが，ここではテキスト形式および HTML でタグ付けされた文献を対象に扱っている．

（1）専門用語

さまざまな分野において，分野に固有の用語が存在するが，それが専門用語（専門語）であり，その分野の特徴を表している．文献に専門用語が数多く出てくる場合には，その文献の特徴を表していると考えられ，専門用語をキーワードとみなして取り出すことを考える[Feldman 98].

（2）複合語

複合語は 2 語以上の語が一組になって一語のように解釈可能であり，それ自体では文になっていない部分であり，句（フレーズ）にあたるものである．複合語にすると意味内容が豊かになり，人間の発想を刺激する効果を高めることができる．そのためここでは複合語もキーワードとして取り上げる方法を取り入れる. 専門用語では複合語が多数を占めるといわれ, 複合語の性質をうまく利用して抽出する. [Hayashi 97]

（3）高頻度語

同じ文献の中で何度も繰り返し使われる高頻度語は，それなりに著者の重要な概念を表していると考えられ，それは重要な部分とみなすべきであり，キーワードとして抽出する必要がある．

日本語処理では，形態素解析によって抽出されたキーワードが検索に使われることがある．形態素解析は，日本語を単語に分割してくれる点で便利である．しかし単語のレベルまで分解されるため，形態素解析の結果をそのまま利用して索引を作成すると，一般的な用語が多くなるため，内容を反映しにくい索引になりかねない.

英語の場合は，言語の性質上はじめから空白で単語が分かれているため，単語を取り出す点ではやりやすいが，意味のある複合語を効率的に取り出すには工夫が必要である.

4. キーワードマイニング

最近は大量のテキストデータから有益なデータを取り出すことをテキストマイニングと呼んでおり，インターネットの普及に伴いますます研究開発が盛んになっている[Fukuda 96, Kawano 01, Nasukawa 99]. ここではキーワードを抽出するので，キーワードマイニングと呼ぶことにする.

マイニングの対象となるデータベースは，インターネットや学術雑誌などから収集した文献である．これらの文献はダウンロードによって収集した段階では，ファイル名以外は検索用の索引を持っていないので，内容を見たいときはそのファイル名からアクセスするのが普通である．Windows のファイル検索や Linux の ls や grep コマンドなどを使えば，ファイル名やファイル内の単語を検索することもできる.

ここではキーワードマイニングによって抽出した専門用語と HTML のリンク機能を使って，これら検索機能の一部を代替し，クリックするだけで内容の閲覧が可能な使いやすい検索インタフェースを持つ機能を実現するためのキーワードマイニングを考えていく[Card 99，Fayyad 02].

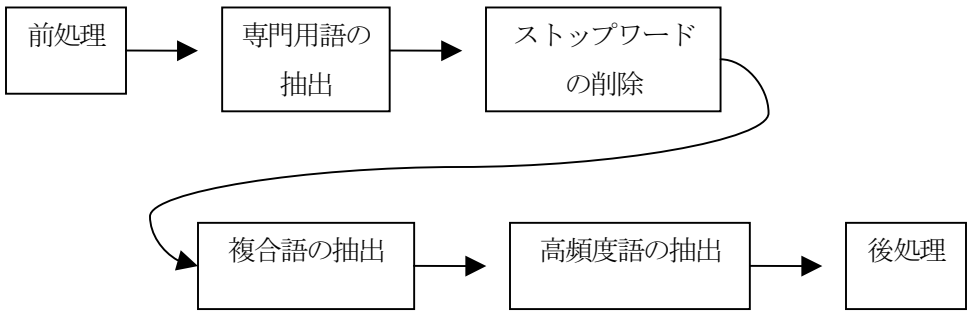

図 3　キーワードマイニングの処理過程

以下では専門用語と高頻度語および複合語を，いくつかの手順を得て段階的に抽出する方法について述べる(図 3)．ここでは扱う分野を地球環境問題として進めるが，専門用語辞典のようなものが存在する分野であれば以下に述べる方法が問題なく応用できる．

またここで提案する方法では，キーワードを抽出する順番が極めて重要であり，順番を変更すると得られる結果も変わる可能性が高い．

処理手順の概要を述べるとまずそれぞれの文献ごとに前処理を行う．前処理では文献のテキストをセンテンスに分割し，HTML タグの削除，単語の接辞処理 (stemming)，ピリオド・カンマの削除などを行う．次に専門用語の抽出，ストップワードの削除，複合語の抽出，高頻度語の抽出を行い，文献ごとのキーワードを取り出す．この処理を集めた文献の数だけ繰り返す．

4．1．　前処理

収集したままの文献から直接キーワードを抽出しようとすると，さまざまな理由から効率が悪いため，事前にいくつかの前処理を行い[Porter 80]，キーワードを取り出しやすくしておく必要があり，英語の文献では次のような処理を行う．

（1）センテンスに分割

ピリオドを頼りにセンテンスに分割し，1 文ごとの処理を行えるようにする．

（2）HTML タグの削除

自然言語で意味を持ち，キーワードになる単語だけを取り出すため，HTML のタグは削除する．

（3）単語の正規化（接辞処理）

専門用語辞書では複数形や現在進行形など変化しているものを原型に戻す．例えば複数形の apples は apple に，進行形の singing は sing に戻される．大文字は略号に多く使われるので，そのまま残している．[Porter 80，Tokunaga 99]

（4）特殊記号の削除

ピリオド，カンマ，括弧などの特殊記号を削除する．

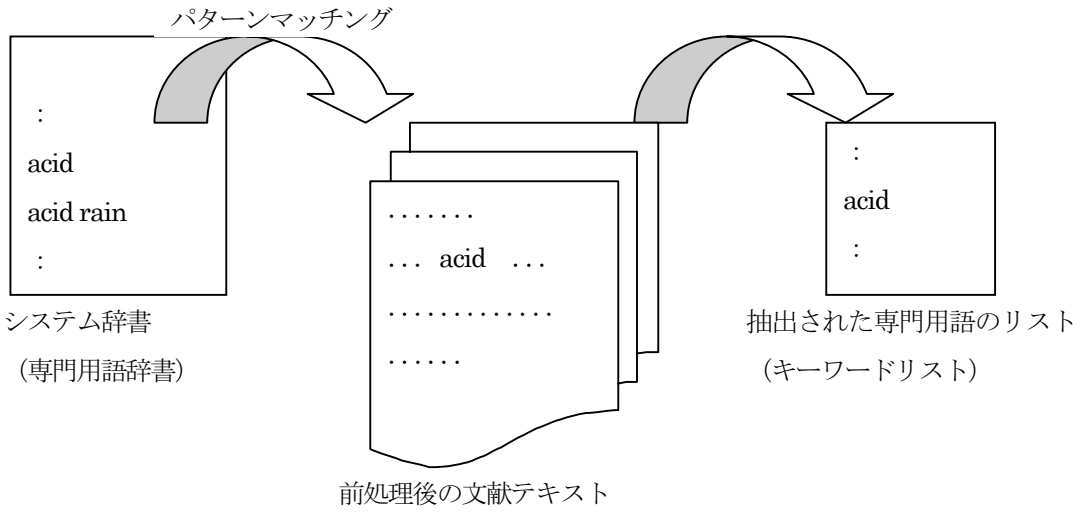

図 4　システム辞書による専門用語の抽出

4. 2.　専門用語のマイニング

コンピュータで文献中のどの用語が専門用語かどうかを判断させるには，最も手っ取り早い方法として，取り出したい専門分野の辞書（専門用語辞書）をシステム上に用意することである．分野によってはインターネットに公開されているものもあり，それらを活用することで個人利用に限定した専門用語辞書を手軽に作成することができる．

ここでは本として市販されている地球環境問題の辞典の索引部分を利用し，それを OCR で読ませてテキストデータを作成し，スペルチェックを行いシステムで使える専門用語の辞書（システム辞書と呼ぶ）として再構成したものである．当初のシステム辞書の用語数は約 1 万語であり，1 単語の専門用語だけでなく複合語や多くの略号を含んでいる．どの分野でもそれぞれ特有の略号が使われているので，それらもシステムに利用する専門用語に含めるべきであり，含めない場合はプログラム上に略号を抽出するアルゴリズムを実装する必要がある．

以下にシステム辞書の一部を示す(図 5)．専門用語は文字数の多い順にソートされているが，この後で作成する HTML の検索用リンクを生成するためである．

実際にマイニングを行う場合は，まずシステム辞書から用語の長い順番に専門用語を 1 つずつ読み込み，パターンマッチングによって前処理したテキストデータ内に一致する部分が出現していないかどうかを調べる．これを全ての処理対照となるテキストデータに対して繰り返し，元の文献ごとに専門用語を取り出す（図 4）．

：　（略）
simulation of the atmospheric dispersion of pollutant
International Council on Monuments and Sites
allowable exposure limit to industrial noise
：　（略）
Universal Declaration of Human Rights
International Association for Ecology
：　（略）
systematic sampling
sea water intrusion
reclamation of land
：　（略）
energy saving
soil bacteria
floral region
：　（略）
hydropower
salt marsh
crown fire
：　（略）
UNESCO
UNICEF
：　（略）

図 5　システム辞書（専門用語辞書）の内容

4. 3.　複合語の抽出

複合語もその文献内容の特徴の一部を表すものと考えられが，出現頻度が 1 回の場合はその複合語を取り出すことは難しい．ここでは出現頻度が 2 回以上の複合語を取り出す方法で対処している．
また複合語を取り出す場合は，できる限り語数の長いものから先に取り出す工夫が必要である．語数の長い複合語の中に専門用語や短い複合語などが含まれているときもあるため，語数の短い複合語を先に取り出すと，語数の長い複合語が破壊されることがあるためである．

これは HTML のリンクを作成する場合に役立ち，長い複合語のリンクの中に専門用語などのリンクを二重に作成することができる．つまりリンクの入れ子構造を作成するために使用可能になる．

以下の例は，gross world product という複合語にリンクが作成され，さらにその中で product にもリンクが生成された例である．以下の例では gross world product をファイル名として使用するため，単語間の空白を_（アンダースコア）に置き換えている．

```
<a href = "/~dobashi/linkfile/gross_world_product.html">
gross world <a href = "/~dobashi/linkfile/product.html"> product </a> </a>
```

複合語のマイニングは特別な辞書は使わず，アルゴリズムだけで行うことができる．4 単語以上の複合語が出現する例は，団体名や組織名などに限られたものになることが多い．そのためここでは 3 単語までの連続した単語から構成される複合語のうち，出現頻度が 2 回以上の部分を句として認識し，抽出することにする．なおプログラム上は 4 単語以上の複合語を抽出することも可能であるが，処理時間もそれなりに必要になる[Ogawa 93].

以下にアルゴリズムの概要を示すが，ここまでの段階では文献に対して専門用語のマイニングを実施していることが前提になっており，専門用語が抽出された部分は文献上では空白に置き換えられており，二重にキーワードを抽出することを防いでいる．

（1）文献の先頭から 3 単語を読み込む.
（2）1 語ずつずらしてパターンマッチングを行い，読み込んだ 3 単語と一致する句が文献内に出現しているかどうか調べる.
（2）もし出現している場合は，頻度をカウントする.
（4）文献の最後までパターンマッチングを行う.
（3）次に文献の先頭から 2 単語を読み込み，同様にパターンマッチングを行い，出現している場合には頻度をカウントする.

例えば次のようにＡＢＣＤ ... というように，単語と単語の間に空白がある文献テキストがあるとすれば，3 単語の複合語はつぎのようにマイニングされる.

（1）ＡＢＣＤＥＦＧ, ＨＩＪＫＬＭＮ, ＯＰＱ.	ABC という先頭の 3 単語を読み込む.
（2）ＡＢＣＤＥＦＧ, ＨＩＪＫＬＭＮ, ＯＰＱ.	1 語ずつずらしてパターンマッチングを行い,
（3）ＡＢＣＤＥＦＧ, ＨＩＪＫＬＭＮ, ＯＰＱ.	ABC の 3 単語と一致するかどうか調べる.

以上を文献の最後まで行い，ABC という 3 単語を複合語の候補と仮定し，これが 2 回以上出現すればそ

れを複合語として取り出す．次は同様にBCDという3単語を複合語の候補と考え，文献の先頭から一致する部分がないか調べる．なお一致する部分があるときは，空白に置き換え，2単語の複合語と重複してマイニングされないようにする．

この考え方は，文献の先頭から3単語（または2単語）を順番に取り出して，複合語の候補となる単語の組み合わせを作成し，その単語の組み合わせが一定回数以上出現したものを複合語とみなすと考えることもできる．

4．4．　高頻度語の抽出

概要などの比較的短い文献では，出現頻度が低くても重要なキーワードになっていることがある．逆に長い論文になると一般的な用語の出現頻度も上がる場合がある．そのため頻度に頼ってキーワードを抽出する場合は，不要語（またはストップワード）を前処理の段階で削除するなどの工夫を行う．

ここでは比較的長い専門用語や複合語の中に不要語が出現する場合があること想定しているため，1単語の高頻度語を抽出する直前に不要語の削除を行っている．

また1単語の高頻度語を取り出す場合は，専門用語と複合語を抽出した残りのテキストデータに対して処理を行う．具体的には空白ごとに単語に分割し，その頻度をカウントするだけである．

キーワードとして抽出する場合は，ここでは2回以上出現した用語をキーワードとして認定しているが，文献の長さを考慮した出現頻度の設定が必要になる．例えば英文で1万単語を超えるような長い文献の場合には，出現頻度を3回以上とするなどの工夫が必要になる場合もありうる．逆に500単語以下の概要文献などは，2回以上の出現頻度でも抽出できるキーワードが数個のこともある．

4．5．　不要語リスト（stop list）

情報検索ではキーワードにはなりえないような用語を不要語あるいはストップワード（stop word）という．「パソコンを買う」という文章があるとき，「パソコン」や「買う」などのように特定の概念を表す語を内容語と呼んでいる．内容語は文献を特徴付けるために有効な働きをしているので，キーワードになりやすいが，「する」や「らしい」などのような語が全部なるわけではない．また「を」のように語と語がどのような関係かを表す機能語の場合には，文献を特徴付けるような有効な働きはないのでキーワードにはならない．

不要語リストはこのようなキーワードとして不適切な語を削除するための辞書である．以下にシステムで使用している不要語リストをアルファベット順に並べたものを示すが（図6），これはSMARTシステムに使われているものを参考に作成したものである[Salton 88]．インターネットにはそれぞれの検索システムで使われているストップワードが公開されているものもあり，それらを見ると採用している語に若干違いがあることがわかる．

この不要語リストを用いて，高頻度語の抽出を行う直前に，リストにある語を削除して空白に置き換え，キーワードとして取り出さないようにする．

a able about above according accordingly across actually after afterwards again against ain't all allow allows almost alone along already also although always am among amongst an and another any anybody anyhow anyone anything anyway anyways anywhere apart appear appreciate appropriate are aren't around as aside ask asking associated at available away awfully b be became because become becomes becoming been before beforehand behind being believe below beside besides best better between beyond both brief but by c c'mon came can can't cannot cant cause causes certain certainly changes clearly co com come comes concerning consequently consider considering contain containing contains corresponding could couldn't course currently d definitely described despite did didn't different do does doesn't doing don't done down downwards during e each edu eg eight either else elsewhere enough entirely especially et etc even ever every everybody everyone everything everywhere ex exactly example except f far few fifth first five followed following follows for former formerly forth four from further furthermore g get gets getting given gives go goes going gone got gotten greetings h had hadn't happens hardly has hasn't have haven't having he he's hello help hence her here here's hereafter hereby herein hereupon hers herself hi him himself his hither hopefully how howbeit however I i'd i'll i'm i've ie if ignored immediate in inasmuch inc indeed indicate indicated indicates inner insofar instead into inward is isn't it it'd it'll it's its itself j just k keep keeps kept know knows known l last lately later latter latterly least less lest let let's like liked likely little look looking looks ltd m mainly many may maybe me mean meanwhile merely might more moreover most mostly much must my myself n name namely nd near nearly necessary need needs neither never nevertheless new next nine no nobody non none nor normally not nothing novel now nowhere o obviously of off often oh ok okay old on once one ones only onto or other others otherwise ought our ours ourselves out outside over overall own p particular particularly per perhaps placed please plus possible presumably probably provides q que quite r rather rd re really reasonably regarding regardless regards relatively respectively right s said same saw say saying says second secondly see seeing seem seemed seeming seems seen self selves sensible sent serious seriously seven several shall she should shouldn't since six so some somebody somehow someone something sometime sometimes somewhat somewhere soon sorry specified specify specifying still sub such sup sure t take taken tell tends th than thank thanks that that's that's the their them themselves then thence there there's thereafter thereby therefore therein thereupon these they they'd they'll they've think third this thorough thoroughly those though three through throughout thru thus to together too took toward towards tried tries truly try trying twice two u un under unfortunately unless unlikely until unto up upon us used useful uses using usually uucp v value various very via viz vs w want wants was wasn't way we we'd we'll we're we've welcome well went were weren't what what's whatever when whence whenever where where's whereas whereby wherein whereupon wherever whether which while whither who who's whoever whole whom whose why will willing wish with within without won't wonder would wouldn't x y yes yet you you'd you'll you're you've your yours yourself yourselves z zero

図 6　不要語リストの内容（ストップワード）

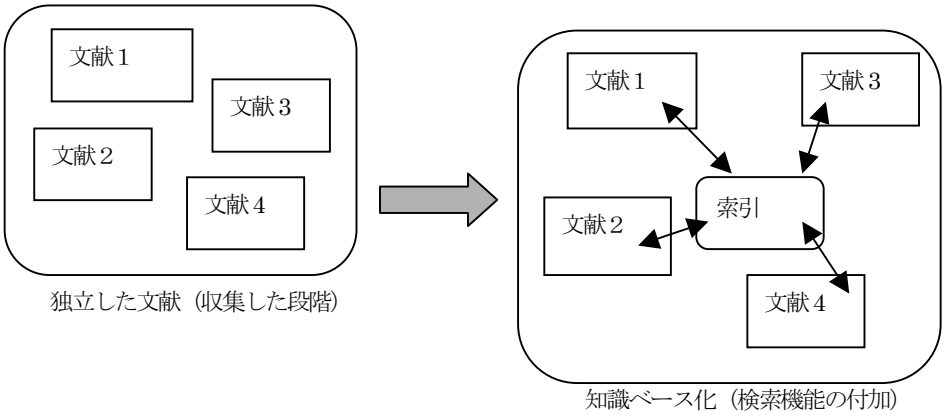

図 7　索引作成と文献の知識ベース化

この不要語リストはテキストファイルなのでエディタを使えば，リストの追加や削除が可能である．文献によっては一般的な用語が極めて高い頻度で出現する場合があるので，そのような用語をキーワードとして採用したくないときは，その語をこのリストに追加すればキーワードとして抽出されなくなる．

例えばWater Resource Research という水資源に関する学術雑誌があるが，この文献からキーワードを抽出すると，water という単語が他のキーワードの頻度よりもはるかに高い頻度で出現する．このような場合は water という一語だけのキーワードは抽出せず，water を含む複合語を抽出したほうがよい場合も考えられる．そのため不要語リストに water を追加すると，water 一語のキーワードは取り出されないが，water と組み合わされた 2 語以上の複合語のキーワードは取り出されるようになる．

5．内容検索への適用

ここまでのキーワードマイニングの処理によって抽出されたキーワードを使い，収集した文献の内容を検索する仕組みを考えていく．検索といってもキーワードを入力して検索するような従来の多くの検索システムで行われているものではなく，リンクをクリックしたら次の内容が見られるというハイパーテキストのリンク機能を基本にするものである．これらのリンク機能による検索は，操作がシンプルではあるが，目的とする情報と関連した情報が得られやすいなどの特徴がある．

5．1．ハイパーテキスト化と知識ベース

収集した文献は定められたあるディレクトリの中に集めてあると想定しており，ここまでの処理で文献ごとのキーワードの抽出を行うことができる．取り出したキーワードは，いくつかの目的に応じて活用することができる．多数の文献が収集されると，それらを検索する機能が必要であることは既に述べたが，キーワードとリンク機能を使えば，クリックするだけで文献をブラウズする機能が実現できる．収集した段階では文献は個々に独立して存在しているが，それらを再構成して検索機能を備えたハイパーテキス

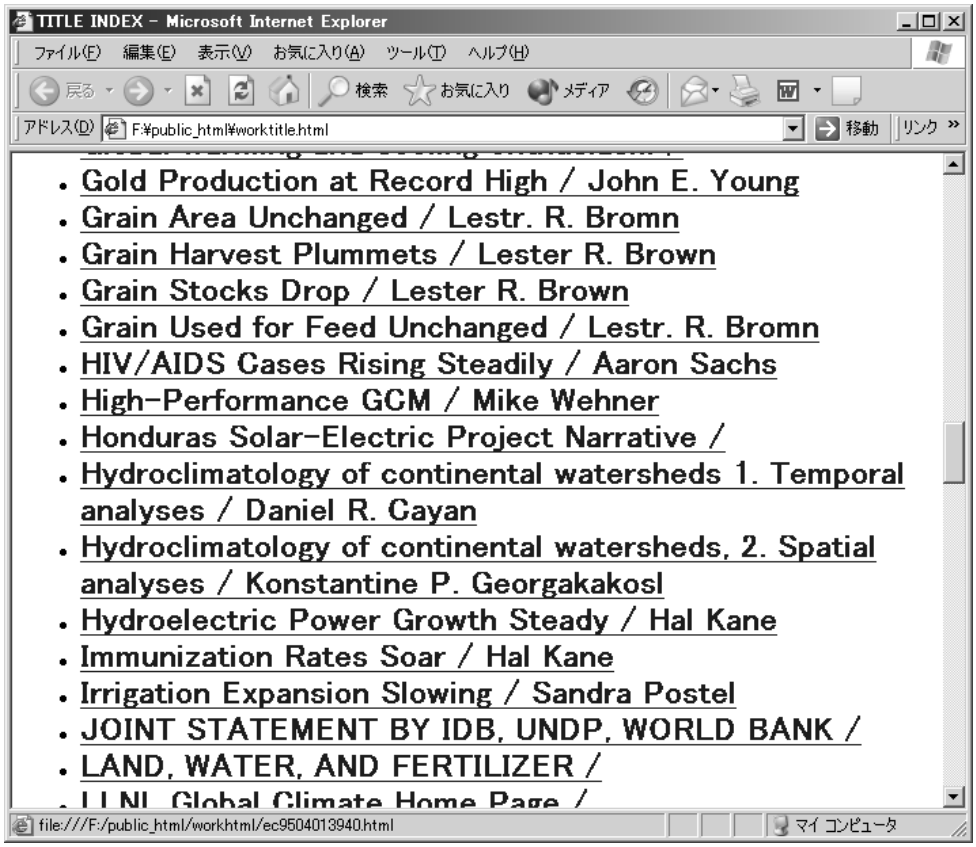

図８　タイトル索引の例

ト化することができる．

また収集した文献が特定のテーマを扱った論文などであれば，ある程度関連性のありそうな知識をあつかっているものと想定され，それぞれの文献ごとに述べられた知識をリンクで関連付けることができる．

すなわち検索機能を持たない無秩序な文献集合に対して，タイトルを一見したところでは内部の関連性が見いだしにくいそれぞれの文献内部に関連性を作り出し，リンクによる検索機能を備えることによって，知識ベースとして再構築することが可能になる（図7）．

インターネットに公開されている非定形で断片的な文書を知識ベース化して体系化すれば，再利用が促進される可能性が高まり，新たな関連性に気づくような効果も期待される[Pratt 99]．しかし現在のインターネットの情報量は既に人間の処理能力をはるかに超えており，これら知的資源の有効活用を支援できる技術が求められている．これは膨大な文献情報が公開される現在のインターネットが抱えている大きな課題でもあり，緊急に解決策の提案が必要とされている問題であり，多くの研究者が関連した研究に取り組んでいる．ここではこのような課題に対して，個人レベルで活用可能なひとつの提案を行っている．

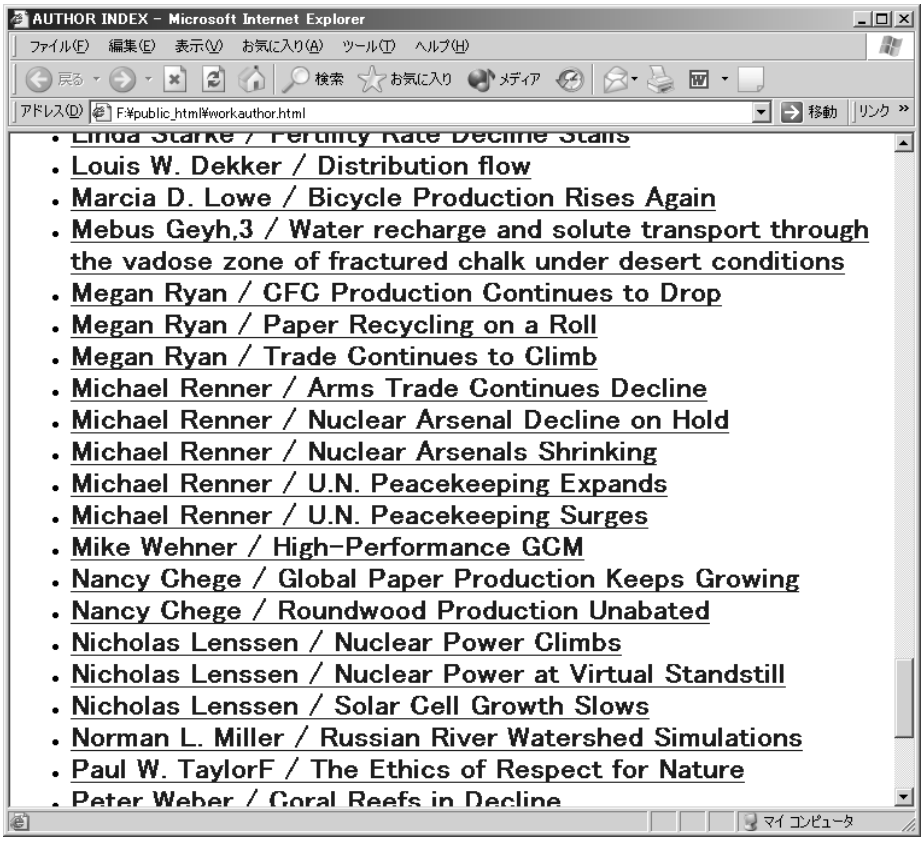

図９　著者索引の例

５．２．　索引の生成

文献検索に必要な索引を作成する場合は，その文献のタイトル名や著者名などがよく使われ，書籍の場合は書名索引や著者名索引がそれぞれ作成される．雑誌記事などの場合は書名の代わりに論文などの題名（タイトル）を集めて論文名の索引（タイトル索引）が作成され，同様に著者名索引も作成される．また最近の書籍の場合には，巻末にキーワード索引が作成されているのが一般的である．

しかし，Web に公開されるページにはこれらのタイトルや著者などの書誌的事項が元々記載されていないものや不適切な場合があり，索引を生成する上での問題となっている．ここでは欠落した書誌的事項を補完するものとしてキーワード索引を生成し検索に活用することを考えていきたい．

そのためタイトルなどは HTML の<title></title>タグに記載があればそれを取り出す．著者についても<meta name="author" content="鈴木一郎">というメタタグか，あるいは<author>鈴木一郎</author>というタグがあればそこから著者名を取り出すことができる．<author></author>タグは HTML ではほとんど見られないが，XML では使われるタグである．ここでは HTML ファイルの作者ではなく，オリジナル文献の著者という位置づけで筆者が独自に使用している．

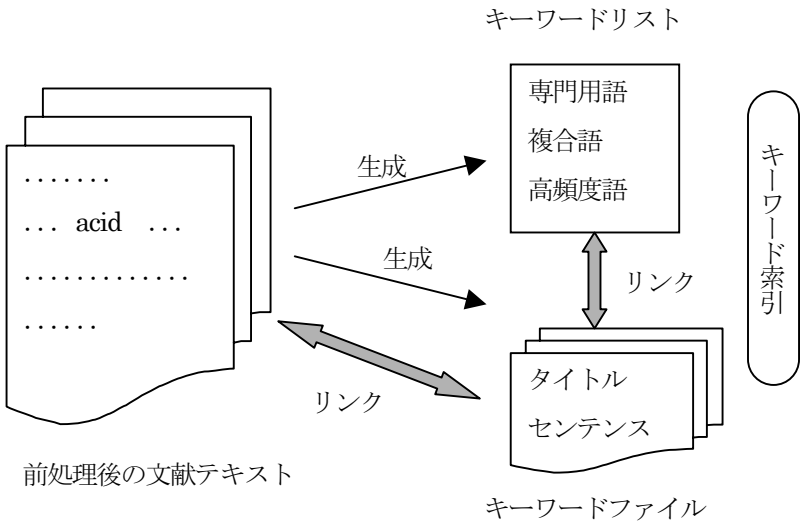

図　10　キーワード索引とリンクの生成

ページによってはこれらのタグが使われずに，タイトル名や著者名が記載されている場合もあるが，このような場合には必要に応じて手作業でタグを作成することにしている．図 8 および図 9 に示したタイトル索引と著者名索引の多くは，筆者が手作業でタグを付与し，そのタグをもとにシステムで生成したものである．タイトルや著者名がない場合は，元の文献にたどり着くための情報が必要になるので，ファイル名あるいは<h1></h1>タグで囲まれた部分や，文献の先頭行を代用するなどの対策が必要である．

5．3．　キーワード索引の構造

キーワード索引は，キーワードリストとキーワードファイルの 2 つの部分から構成される(図 10)．キーワードリストは，抽出した全部のキーワードをアルファベット順に並べた 1 つのファイルである（図 12)．キーワードファイルは，そのキーワードがどの文献のどのセンテンスに出現するかを示すファイルで，抽出されたキーワードごとに作成されるものである（図 13)．

キーワードリストのそれぞれの用語と，キーワードファイルは HTML のリンクで連結しており，またキーワードファイルは文献とリンクしており，キーワードから文献を探し出すために利用することができる（図 13 および図 14)．

キーワードファイルは，キーワードが出現するセンテンスとその文献のタイトルをインデックス化した 2 つの部分から構成される．リンクは共起するキーワード間およびタイトルや著者名と該当する文献間に生成される（図 14)．

これらの索引と知識ベースの間は，リンクによって自動的に階層化されており，利用者が検索しやすいように，知識ベースの内容を視覚化する工夫をおこなっている[Watanabe 01]．またキーワードファイルの中では，キーワード（可能な場合はタイトル名や著者名も含めて）が出現するセンテンスを表示しており，

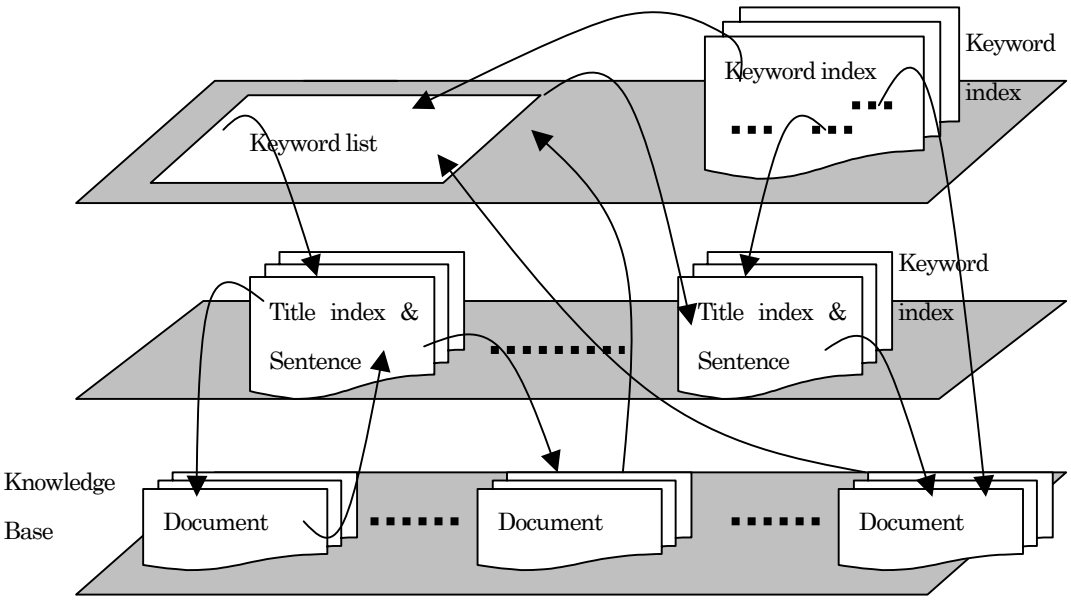

図　11　索引と知識ベース間の階層関係（概念図）

この索引をブラウジングすることによって，利用者の検索目的に関連の深いタイトルやセンテンスが自然に目に入るようになっている（図 14）.

これは情報検索の観点から，利用者の検索要求に関連した部分を表示することによって，問題の関連構造を広く把握することがねらいである[Misue 99]. 情報検索では，必要な情報がダイレクトに得られればよい場合と，そうでない場合も必要なことがある．検索に対する要求があいまいなときは，情報検索の過程で接する関連したいろいろな情報が，利用者にとって役立つ場合がある.

ここではキーワードリストやキーワードファイルを元の文献とは別に作成して検索機能を実現しているが, HTML を元の文献に埋め込んで検索機能を実現する方法もある. 文献内のキーワードの直後や右肩に，そのキーワードのリンク先の情報を表示し，そこからキーワードが出現する文献へ直接リンクすることも可能である．しかしこの方法では，リンク先が多数になると文献上の表示がわずらわしくなり，多数の文献を閲覧するときに起こりやすいハイパーテキスト特有の迷子状態に陥りやすくなる.

5. 4.　キーワード索引の目的

情報検索においては，ユーザが情報を利用する動機のようなものが存在する．ユーザが情報を利用しようとするのは，情報を必要とする何らかの問題を抱えているからであり，その問題を解決したいためである．このような情報要求の状態はユーザによって異なるものであるし，同じユーザでも調べたいことがらについての知識の深さによって異なるものである.

キーワード索引は，できる限り一般的な用語を排除し，専門用語としての意味を把握しやすいように抽

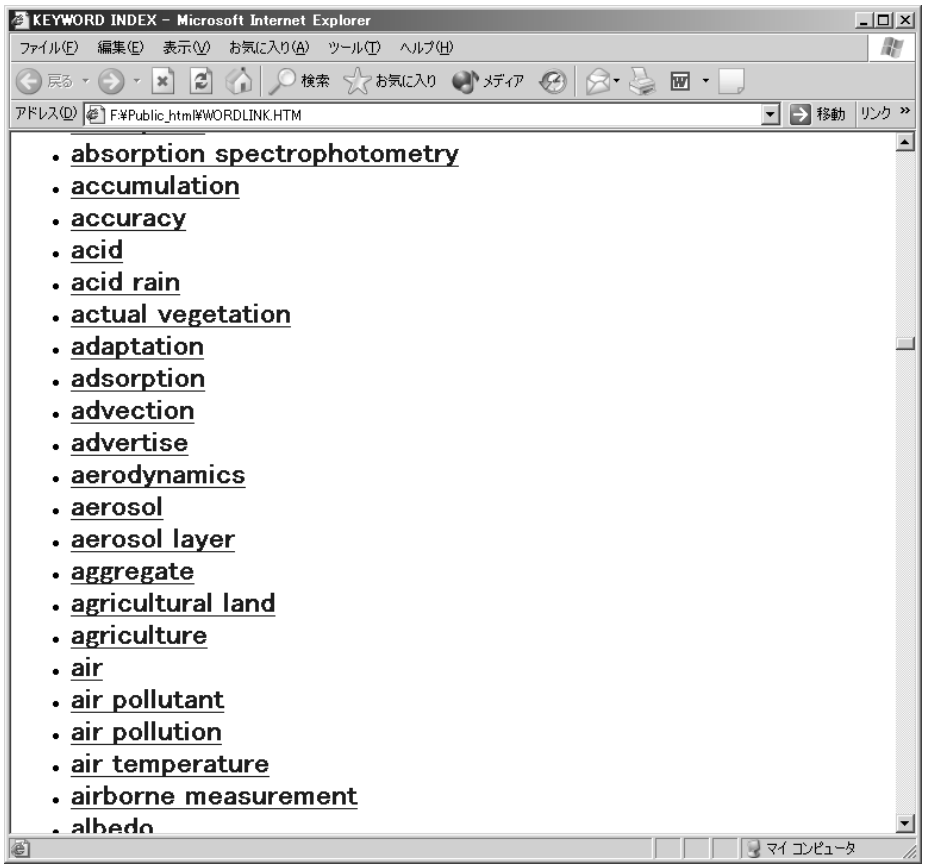

図　12　キーワードリスト

出しており，ユーザの次のような段階に対応することを目標に作成している[Tokunaga 99].

（1）直観的要求の段階

必要な情報への要求が，具体的に言語化して説明できないような状態のときであり，このようなときはキーワードリストやタイトル索引などの索引をブラウズして，要求の言語化を支援する.

（2）意識された要求の段階

あいまいな表現やまとまりのない表現でしか言語化できないような状態のときは，上と同様に索引などをブラウズすることによって実際の言葉に触れ，要求が言語として具体化するように支援する.

（3）形式化された要求の段階

情報への要求が具体的な言葉で言語化できる状態のときは，キーワードリストやタイトル索引などの必要な部分を直接クリックして検索する.

（4）調整済みの要求

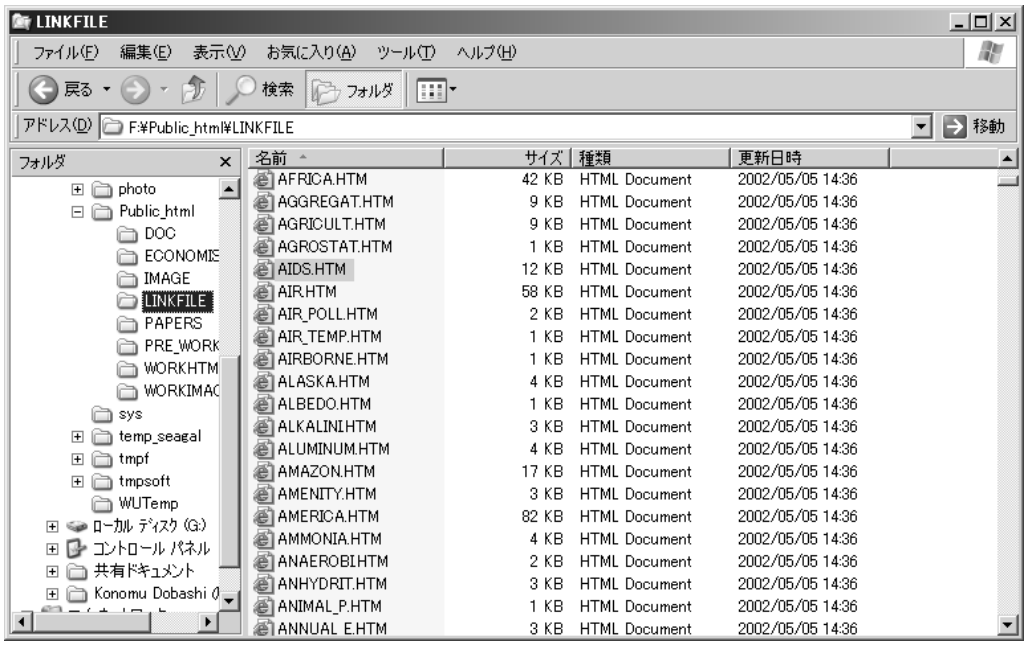

図 13　キーワードファイル

情報への要求を満たすために必要な情報源が同定できるぐらいに問題が具体化された状態のときは，キーワード索引を使ってそのキーワードが出現するセンテンスを検索することができる．

5. 5.　新たなキーワードの追加

ここでは専門用語の抽出に辞書を用いているため，常に同じ状態の辞書を用いて抽出処理を行っていると，追加した文献に新たな専門用語が含まれている場合に取り出すことができないことも起きる．システム辞書に登録されていない場合でも，複合語や高頻度語として抽出されたときは，それらを新しいキーワードとしてシステム辞書に追加することにしている．この処理は新たな文献が追加されたときに行うもので，文献が追加されるたびに複合語と高頻度語を，新しいキーワードとしてシステム辞書に追加する．

しかしこの方法でも，単語の出現頻度が 1 回のときや，複合語にもなっていない専門用語は抽出が難しく，一般的な用語を含めるとノイズとなるので工夫が必要である．ユーザが索引にない専門用語に気付いたときは，手作業ではあるがシステム辞書を編集して追加できる．

6.　関連研究

近年，多くの文書が電子化されて蓄積されるようになり，WWW の普及によりそれらの文書が大量に流通するようになり，手軽に入手できるようになってきた．膨大な文書の中から必要な情報を探したり，蓄積された文書を分析して何らかの傾向をつかみたいというような，利用者のさまざまな要求に対応するた

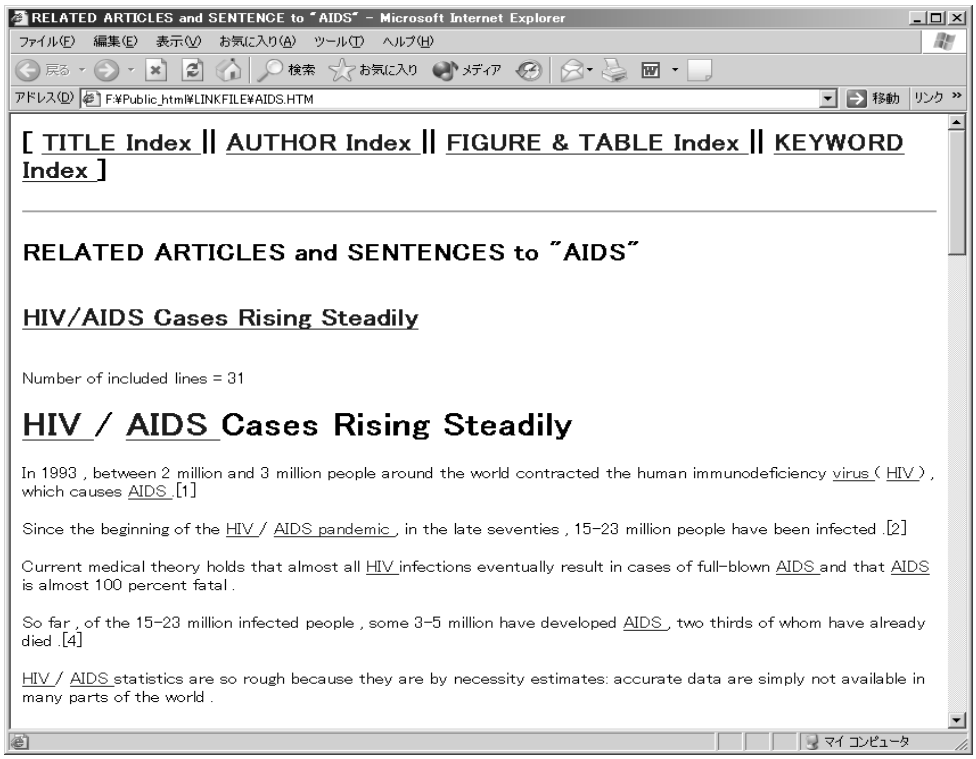

図　14　キーワードファイルの内容

めに, テキストマイニングを活用しようという研究が盛んである[Fukuda 96, Nasukawa 99, Ichimura 01].

テキストマイニングは，自然言語処理や情報検索などさまざまな技術を組み合わせた複合技術である[Nasukawa 01a]．なかでも文書処理の基盤となる自然言語処理技術[Tsujii 02]，処理結果などを統計的に分析する統計解析，分析結果をユーザに分かりやすく視覚化するインタフェース技術が重要なものになっている.

自然言語処理では，形態素解析や構文解析を活用してテキストから重要語を抽出したり，文書を自動分類したりすることが行われている．処理結果の統計解析では，相関分析やクラスター分析，数量化理論などが使われる．また情報の可視化技術を取り入れて，単語の連想関係を視覚的に表現する工夫も行われており，システムをインタラクティブに操作し，さまざまな発想効果をねらった研究も行われている[Misue 99, Watanabe 99]．これらの技術のいくつかは組み合わされてシステムとして実現され，最近では企業におけるコールセンターの顧客対応に活用されたり[Nasukawa 01b]，また社内の文書管理や活用を目的に商品化されているマイニングシステムもある.

さらにテキストマイニングはWebページの検索をより高速化する研究や, Webページの検索結果から有用な情報を見出す研究などにも応用されている[Abe 00，Kawano 01，Sakamoto 01].

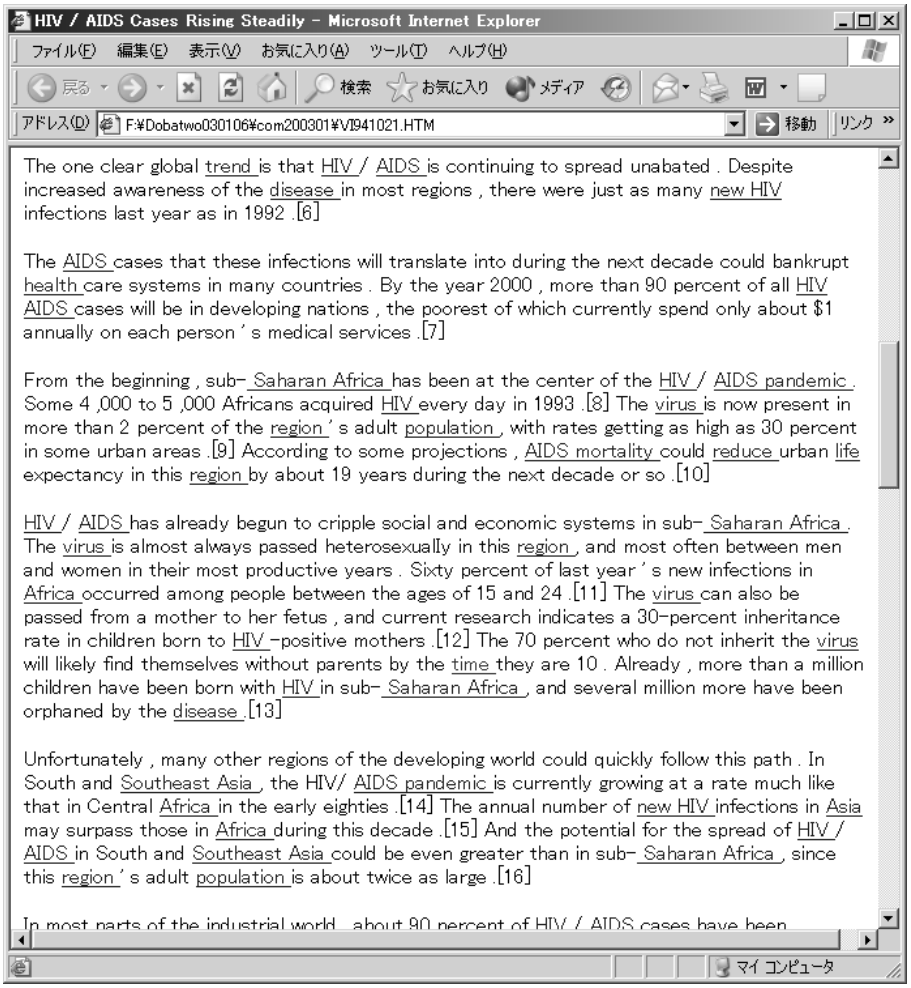

図 15　文献テキストと内部に生成されたリンク（下線部分）

（Aaron Sachs 著，HIV / AIDS Cases Rising Steadily，Vital Signs 1994 (pp.102-103)より作成）

7. まとめ

ユーザの情報要求は，検索の途中で新しい情報に出会うことによって，動的に変化することが考えられるため，これに対応する検索の仕組みが必要である．インターネットの Web の検索を行うときが，まさにこのような情報要求の変化が起きやすいといえる．このようなときは情報を求めてキーワードを入力するような検索方法よりは，あちこち拾い読みできるような機能が役に立つ．

キーワード索引はこのようなときに有効に機能するように索引の用語を限定し，検索機能を付加したつ

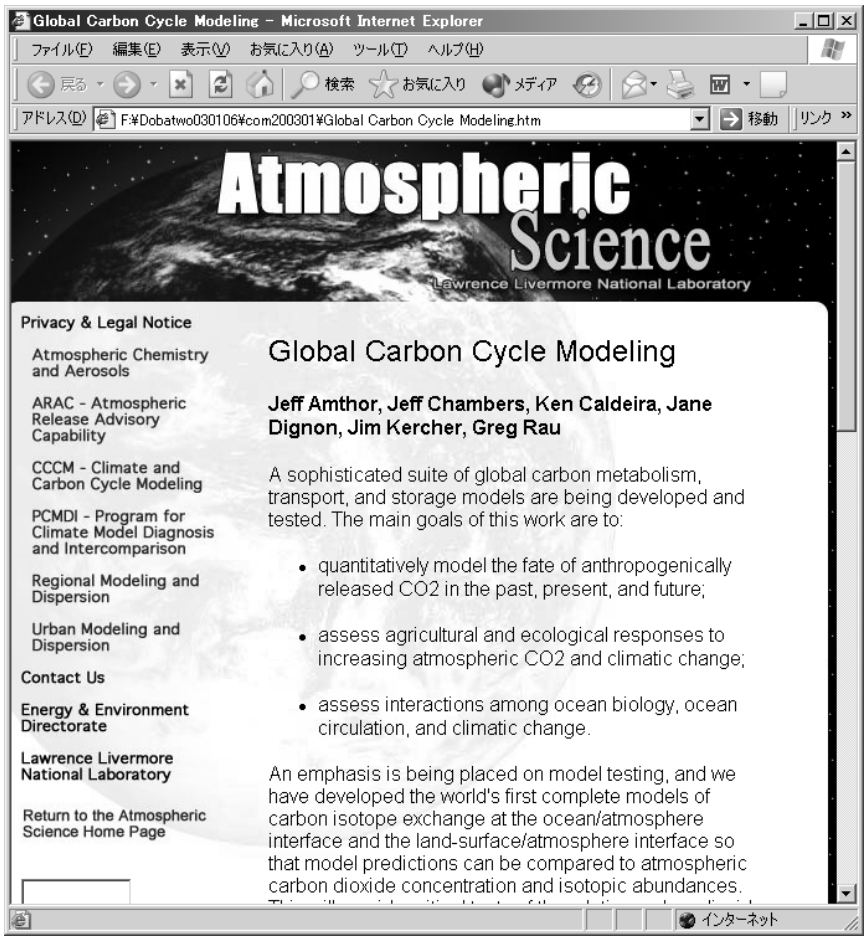

図 16　Web からダウンロードしたページ（http://en-env.llnl.gov/asd/c-cycle.html）

もりであるが，キーワードを入力する従来型の検索機能を備えていないなど，現状では全ての要望に応じられるほど十分ではない．しかし従来型検索機能との統合は難しいことではなく，また新たな機能の追加も視野に入れて考えれば，今後さらに発展させることができそうである．

また専門用語を抽出するときに使うシステム辞書の性能が，キーワードの抽出に大きく影響する．さらに学際的な分野で情報を集めようとすると，複数の分野にまたがるシステム辞書が必要になる．実用に耐えるシステム辞書を効率的に構築する支援機能の充実と，辞書を使用しないでキーワードを抽出する方法の採用などの検討が必要である．

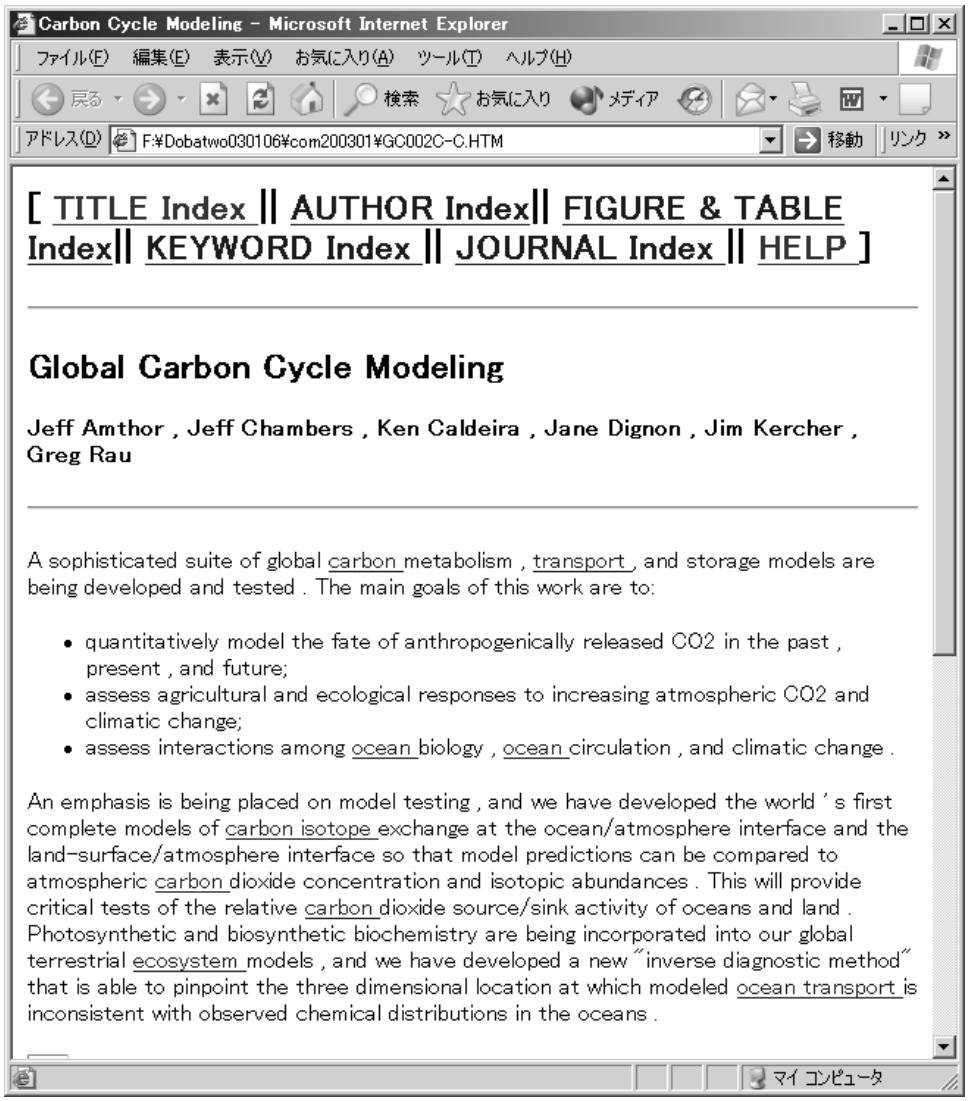

図 17　図 16 のテキスト部分に作成したリンク（下線部分）

謝　辞

本研究をまとめるにあたり愛知大学研究助成を受けた．ここに感謝の意を表す．

参考文献

[Abe 00] 安部潤一郎, 藤野亮一, 下薗真一, 有村博紀, 有川節夫, "テキストデータからの高速データマイニング-探索的文書ブラウジングとウェブデータへの応用-"，人工知能学会誌, 特集「発見科学」，Vol.15,

No.4,pp.618-628(2000).

[Card 99] Card, S. K., Mackinlay, J.D., Shneiderman, B., "Readings in Information Visualization Using Vision to Think", Morgan Kaufmann, pp.686 (1999).

[Fayyad 02] Fayyad. U. M., Grinstein G.G., Wierse A., "Information Visualization in Data Mining and Knowledge Discovery", Morgan Kaufmann, pp.407 (2002).

[Feldman 98] Feldman, R. et al. "Text Mining at the Term Level", In Proc. of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98), pp. 65-73(1998). http://citeseer.nj.nec.com/feldman98text.html

[Fukuda 96] 福田剛志, "データマイニングの最新情報－巨大データからの知識発見技術－", 情報処理, Vol. 37, No. 7, pp.597-603 (1996).

[Hayashi 97] 林淑隆, 中野英雄, 獅々堀正幹, 青江順一, "文字列照合マシンを利用した複合語キーワードの効率的抽出法", 情報処理学会論文誌, Vol.38, No.4, pp.815-825(1997).

[Ichimura 01] 市村由美, 長谷川隆明, 渡部勇, 佐藤光弘, "テキストマイニング－事例紹介", 人工知能学会誌, 特集「テキストマイニング」, Vol.16, No.2, pp.192-200 (2001).

[JEITA 00] 電子情報技術産業協会（編）, "自然言語処理システムに関する調査報告書", pp.183-216(2000).

[Kawano 01] 河野浩之, 川原稔, "Web 検索におけるテキストマイニング", 人工知能学会誌, 特集「テキストマイニング」, Vol.16, No.2, pp.212-218 (2001).

[Nagao 76] 長尾真, 水谷幹男, 池田浩之, "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).

[Nasukawa 99] 那須川哲哉, 諸橋正幸, 長野徹, "テキストマイニング－膨大な文書データの自動分析による知識発見－", 情報処理, Vol.40, No.4, pp.358-364 (1999).

[Nasukawa 01a] 那須川哲哉, 河野浩之, 有村博紀, "テキストマイニングの基盤技術", 人工知能学会誌, 特集「テキストマイニング」, Vol.16, No.2, pp.201-211 (2001).

[Nasukawa 01b] 那須川哲哉, "コールセンターにおけるテキストマイニング", 人工知能学会誌, 特集「テキストマイニング」, Vol.16, No.2, pp.219-225 (2001).

[Ogawa 93] 小川泰嗣, 望主雅子, 別所礼子, "複合語キーワードの自動抽出法", 情報処理学会, 研究会報告「自然言語処理」(93-NL-97), pp. 103-110(1993).

[Ohsawa 99] 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, No.2, pp.391-400 (1999).

[Porter 80] Porter, M.F. "An algorithm for suffix stripping" Program, Vol.14, Num.3, pp.130-137(1980). http://www.cs.jmu.edu/common/projects/Stemming/

[Pratt 99] Pratt, W., Hearst, M., and Fagan, L.; "A Knowledge-Based Approach to Organizing Retrieved Documents;" AAAI-99: Proceedings of the Sixteenth National Conference on Artificial Intelligence. July 1999. http://www.ics.uci.edu/~pratt/pubs/AAAI-99.pdf

[Rajman 97] Rajman M., Besançon R.. "Text Mining: Natural Language techniques and Text Mining applications". Proc. of the 7th IFIP 2.6 Working Conference on Database Semantics (DS-7) (1997).

[Sakamoto 01] 坂本比呂志, 有村博紀, "Web マイニング", 人工知能学会誌, 特集「テキストマイニング」, Vol.16, No.2, pp.233-238 (2001).

[Salton 88] Salton, G., "Automatic Text Processing", Addison-Wesley, pp.530 (1988).

[Tokunaga 99] 徳永健伸, "情報検索と言語処理", 東京大学出版会, pp.234 (1999).

[Tsujii 02] 辻井潤一, "ゲノム情報学と言語処理", 情報処理, Vol.43, No.1, pp.36-41 (2002).

[Misue 99] 三末和男, 渡部勇, "テキストマイニングのための連想関係の可視化技術", 情報処理学会研究報告 99-FI-55,99-DD-19, 情処研報, Vol.99, No.57, pp.65-72 (1999).

[Watanabe 99] 渡部勇, 三末和男, "単語の連想関係によるテキストマイニング", 情報処理学会研究報告 99-FI-55,99-DD-19, 情処研報, Vol.99, No.57, pp.57-64 (1999).

[Watanabe 01] 渡部勇, "ビジュアルテキストマイニング", 人工知能学会誌, 特集「テキストマイニング」, Vol.16, No.2, pp.226-232 (2001).