人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第10回)
SIG-AM-10-01

文書構造に基づく対話的情報アクセスにむけて

Towards Interactive Information Access based on Document Structures

加藤 恒昭[1*]　岩月 憲一[1]　山口 和紀[1]
Tsuneaki Kato[1]　Kenichi Iwatsuki[1]　Kazunori Yamaguchi[1]

[1] 東京大学 大学院 総合文化研究科
[1] The University of Tokyo Graduate School of Arts and Sciences

Abstract: A framework is examined, in which the users interactively access documents, like scientific papers, with a physical structure appearing in the layout and a logical structure based on their contents. It supports effective and flexible use of the documents by allowing the users to retrieve relevant logical units through specification of their contents and/or roles in the document, and to browse those units and their contexts by strolling across both logical and physical structures. The whole framework and a method of document analysis that reconstructs the logical structure of a document and constructs its representation are mainly discussed in this paper.

1　はじめに

一般に文書は，章立てのような意味内容に基づく論理構造と，印刷・表示される場合のレイアウトに対応する物理構造を持つ．本稿では，これらの構造を利用することで，様々な検索意図に対応しうる情報アクセス環境が構築できることを述べる．まず，情報アクセスにおいて，文書全体でなく，文書の構造を用いてその部分にアクセスできることの必要性を述べ，そのような構造が対話的な情報アクセスにおいても重要であることを指摘する（2 節）．続けて，文書構造に基づく情報アクセスによってどのような検索意図に応えられるかを掘り下げ，そのために必要な文書表現を検討する（3 節）．その後，そのような文書表現を得るための文書の論理構造抽出について，方針と現状を報告する（4 節）．最後に関連研究について言及し（5 節），今後の方針を述べて全体をまとめる（6 節）．

以下，学術論文や学会発表予稿集，特に言語処理学会 20 周年記念で公開された年次大会予稿集[1]を，構造を持つ文書の例として議論を進めるが，その議論は，意味内容に基づく論理構造と，それと結びついたレイアウト等の物理構造を持つ情報源に自然に拡張できる．例えば，Wikipedia のようなマルチメディア事典，コマ割りという論理構造かつ物理構造を持つコミック等についても，同じようなニーズが存在し，同じ枠組みで捉えることができると考えている．

*連絡先：東京大学大学院総合文化研究科言語情報科学
〒 153-8902 東京都目黒区駒場 3-8-1
E-mail: kato@boz.c.u-tokyo.ac.jp

[1] http://www.anlp.jp/resource/annual_meeting.html

2　情報アクセスと文書構造

一般に文書として流通している情報は，情報アクセスの単位として必ずしも適当なものでなく，文書の構成要素に直接アクセスできることが必要である．例えば，学術論文や学会発表予稿集は研究活動を進めるにあたっての重要な情報であり，様々な検索意図に基づいた情報アクセスが行われる．それらに答えるために必ずしも文書全体が必要なわけではない．ある評価指標の定義が知りたいのであればひとつの式がその回答になるであろうし，その評価指標を利用するための評価実験の概要が知りたければ，論文の一節だけを提示すればよい．その評価指標がどの程度一般的なものであるかを知りたいのであれば，それを用いている論文の数だけでも参考になる．この例のような文書の一部に関心があるという場合に限らず，そこで述べられている研究そのものに興味関心がある場合でも，利用者は論文を最初から丁寧に通読していくわけではない [16]．梗概や導入だけを読んで，その価値を，読み進めるに値するかを判断することも多い．であればまずはその部分だけを提示するのが適切であろう．

文書全体ではなくそこに含まれる特定の情報が利用者のニーズを満たすということは，パッセージ検索 [4, 6, 12] や質問応答 [15] の動機となっている．ただ，初期のパッセージ検索の動機は文書の適合性を測る場合にそれ全体の特徴ではなく，その部分に注目した方がよいというものであるし，質問応答は文書全体の主題と無関係にそこに含まれる情報を利用しようというものであった．そこでは，文書の構成要素が文書とは独

立に扱われていて，構成要素が文書という構造の中である役割を持っており，それに基づいてアクセスされるという視点は弱い．上述の評価や梗概の例のように，文書の構成要素はそれ自身の特徴だけでなく，文書という構造の中での役割に基づいて利用できることが求められる．あわせて，これらの取り組みでは，対話的な情報アクセスの観点が欠けている．

学術論文を含め，様々な情報の活用は対話的・探索的に行われる．複数の検索結果を斜め読み的に閲覧して，必要な情報を見定めるという，既に述べたような利用に加えて，ある評価指標の定義からその利用方法への関心の拡大．関心を持った文書からそこで引用されている文書への推移等，Bates のいう Berrypicking[2] での推移，Ellis のモデルにおける Chaining[5] のような推移に対応しなければならない．文書間の推移については，例えば文書を引用関係で結び付けたハイパーテキスト構造を閲覧の対象とすること等が試みられているが，文書内に閉じた閲覧やブラウジングにおいても，それぞれの情報の文脈を提示することや概要から詳細への焦点の推移が重要になる．最初の例に戻れば，評価指標の式からそれを含んだ評価実験の記述への推移や，その逆の推移が自然に行えることが望ましい．その点でも，文書を単位とせず，文書の構造を意識することが必要である．そして，そのような文脈や構造を利用者に自然に提示するものとして，論文誌，予稿集に掲載されていてレイアウト，物理構造が有益であることが期待される．このような形式は文書閲覧の形式として馴染みがあることに加えて，一般にはテキスト検索の対象とならない図表類を情報として含んでおり，対話的な検索を通じてそれらの情報を提供する機会を与えることになる．

このような着眼に基づいて，1) 文書を意味内容に基づく論理構造を持つものと捉え，情報アクセスの単位をその構造の構成要素とするような情報アクセス環境の実現を検討する．論文等の場合，文書の論理構造はいわゆる章立てに対応し，あわせて，タイトルや著者情報，参考文献などが論理構造の構成要素（論理要素）となる．ここで，単に文書を小さな単位に分割・分解するのではなく，それぞれがどのような文脈にあったか，どのような構造の一部であったか，を維持し，検索意図との照合やその後のインタラクションに利用する．2) このような情報アクセスを対話的プロセスの一部とするために，文書が論理構造のみでなく，レイアウトのような物理構造を持ち，図表等の視覚情報を含むことを活かした閲覧やブラウジング等のインタラクションを検討する．レイアウト等の物理構造は論理構造と一定の関係を持つが，必ずしも同じものではない．検索が論理構造に基づいて行われるので，このようなインタラクションはあわせてこの論理構造を意識し，物理構造と論理構造を行き来できなければならない．

3 検索意図との照合

前節で述べた様々な検索意図について分類し，それに応えるためにどのような情報が必要かを検討する．

検索意図は，まず，文書（この場合は研究論文）そのものを必要するものとその部分（構成要素）で応えられるものとに分類される．研究論文はすべて何らかの研究について論じていると看做せるので，その研究を特徴付ける概念が，文書の主題となる．したがって，文書そのものへの検索意図は研究に関する記述を求めていると考えられるが，その研究の指定の仕方は大きく以下の３つに分けられる．

1. 主題に基づくもの
   例:「WordNet についての研究」
2. その他の情報によるもの
   例:「知識源として WordNet を用いている研究」
3. メタ情報（書誌情報）によるもの
   例:「2014 年以降に発表された研究」

知識源や評価尺度として何を利用しているか，どのような文献を参照しているか等は必ずしも主題として研究を特徴づけるものではないので，1. と 2. は区別される．著者や著者が所属する組織等文書そのものから得ることができるメタ情報もあるが，情報とメタ情報の違いとして 2. と 3. が区別される．2. の検索意図に応えるためには，文書の主題を反映する文書表現だけでなく，特定の役割や部分における特徴を蓄積する必要がある．典型的な例は参照している文献による研究の検索で，文書の参考文献の部分に指定された文献が含まれることが条件となる．

一方，文書の部分，その構成要素に対する検索意図は，文書を介するか否かで分類できる．文書を介さない検索意図は，あるキーワード，例えば，WordNet や相互情報量の定義や説明を知りたいというようなもので，その回答はどのような研究で使われているかに関係しない．これは質問応答技術が扱うような検索意図に近く，文書の構成要素毎にその特徴を表現し，適合するものを選択し，更に必要に応じてその一部を抽出して回答することが求められる．一方，文書を介するものは，前述のいずれかの方法で研究を指定し，それに関連する情報を求める．「〜研究における評価手法を知りたい」「〜研究においてよく参照される文献を知りたい」が例となる．この場合，それが文書に対して持つ役割に基づいて，構成要素が検索意図に適合するかを判断する必要がある．例えば，ある構成要素がその研究の評価手法についての部分であることが表現されていなければならない．

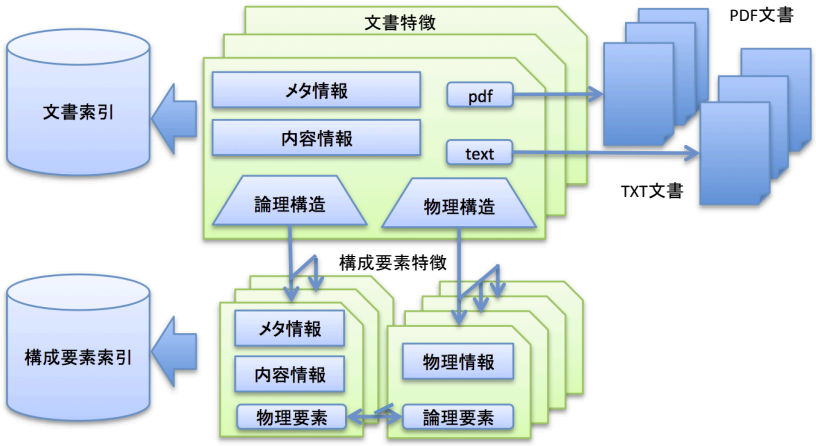

図 1: 文書の表現

このような様々な検索意図に対応するためには少なくともふたつのことが必要となる．ひとつは，表現された検索要求の背後にある検索意図の曖昧性の解消あるいは，その広がり（diversity）に配慮した検索方針で，例えば,「WordNet」という要求で表されている意図が,「WordNet についての研究」「WordNet を使った研究」「WordNet とは何か」等のいずれであるかを明らかにする必要がある．同様に「統計的機械翻訳の評価」は,「統計的機械翻訳の評価についての研究そのもの」や「統計的機械翻訳についての研究の評価」を求めている場合がある．

もうひとつは，そのような意図を満たすための文書表現と照合方式で，上で述べたように，文書の主題に関する表現だけでなく，メタ情報や，その構成要素に関する情報が必要となる．構成要素に関する情報としては，その主題に関する表現に加えて，文書における役割が明らかにされている必要がある．この役割情報は構成要素のメタ情報であり，それによって，文書を選択する条件に関連する部分であるかや，文書中の求められている部分であるかが判断される．これらを適切に使い分けて検索意図との照合を行う必要がある．

このような照合とその後の閲覧を考えた場合に，蓄積すべき文書表現と関連情報を図 1 に示す．文書はそのレイアウトを維持した PDF 文書とそこに含まれるテキストを抽出した TXT 文書として記憶され，そこから取り出された様々な情報が文書特徴として記述される．その中にその論理構造と物理構造の記述がある．論理構造と物理構造は対応づけられ，論理構造のそれぞれの要素については，そこに含まれるテキストについての内容情報と文書中での役割を示すメタ情報が記述され，物理構造の要素にはレイアウトにおける位置情報等が記述される．次節で述べるが，物理構造の要素（基本要素と呼ぶ）は論理構造と n:1 の対応を持つ．これらの文書特徴，構成要素特徴から検索に用いられる索引情報が生成される．

4 論理構造の抽出

4.1 方針

前節で述べた文書表現を獲得するために，文書からその物理構造と論理構造を抽出する検討を進めている．文書として予稿集等の PDF 文書を想定する．PDF 文書は LaTeX や MSWord 等の文書作成組版システムによって直接作成されるディジタル文書と紙媒体の文書をスキャンして得られるスキャン文書に分類される．言語処理学会年次大会予稿集においては，2003 年まではスキャン文書，それ以降はディジタル文書となっている．

スキャン文書から検索可能なテキスト情報と物理構造および論理構造を抽出するためには，OCR ソフトウェアを用いる．一般に OCR 処理はレイアウト認識と文字認識からなる．レイアウト認識は文書の各ページを矩形領域に分割した後，それらをテキスト，表，図等に分類し，位置や大きさの情報を得る．その後，テキストと分類された矩形領域を単位として，そこに含まれる文字の文字認識が行われ，テキスト情報が抽出される．日本語文書の OCR ソフトウェアにおいては，e-typist[2] とその上位製品である Win Reader Pro[3] が，認識結果を xhtml 形式で出力する機能を持ち，そこでは認識された矩形領域が xhtml の span 要素と対応し，その属性として，矩形の位置や大きさが表現される．OCR ソフトウェアのレイアウト認識と文字認識は，ともに完璧ではない．レイアウト認識の問題は後述するが，文字認識においても，特にスキャンの質が低い文書では誤

[2] http://mediadrive.jp/products/et/
[3] http://mediadrive.jp/products/wrp/index.html

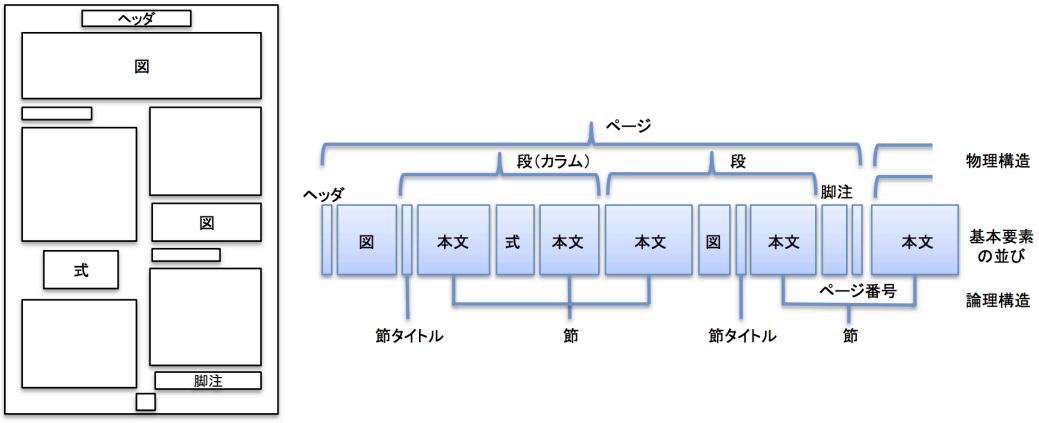

図 2: 論文のページレイアウトと物理構造と論理構造

りが多発するし，数式に使われるような記号としてのアルファベットは殆ど扱えない．このため，OCR 処理には人手介入が許されており，文字認識結果の後修正だけでなく，レイアウト認識を人手で修正した後に文字認識を行うことも可能となっている．

ディジタル文書は，その内部にテキスト情報を持っており，pdftotext[4]などのソフトウェアでこれを抽出することができる．この場合，抽出結果に OCR ソフトウェアの文字認識で生じるような誤りはない（ただし，[7]）．一方で，ほぼ行単位で抽出される文字列の順序は必ずしも文書作成者が意図したあるいは一般的な読者が読み進む順序とは一致しない．また，文字の位置についての情報は得ることができるが，OCR ソフトウェアのレイアウト認識で得られるような人間の直観にあった矩形領域への分割は取得できない．ディジタル文書を html 等に変換するものも配置されるのは行であり，OCR ソフトウェアのレイアウト認識における矩形のような概念は存在しない[5]．

OCR ソフトウェアのレイアウト認識は空白部分の存在（スペーシング）等の情報を用いて矩形領域を認識する．それらは文書の論理構造や意味内容を意識していない．一方，前節で述べた目的のためには，物理構造は論理構造と一定の関係をもつ必要がある．具体的には，論理構造の単位となるものが，紙面の物理的な制約の下で必要に応じて分割され，配置された構造を物理構造と考える．物理的な制約とは，多段組みにおける段の境界，ページの境界，図の挿入，脚注の挿入，ヘッダやフッタの存在などである．例えば，図 2 において，図の左に概念的に示すような論文の 1 ページについて，矩形で囲った部分それぞれを物理構造の基本要素と考える．これらの要素は 2 次元的に配置されているが，２段組の原稿であることを考慮すると，簡単な規則によって図の右に示す 1 次元の並びとすることができる．物理構造を考えた場合，並べられた基本要素が，段やページ等を構成していくし，論理構造を考えた場合は，節やそのタイトル等の物理要素が得られる．物理構造においては常に連続した要素がより大きな構造をなしていくが，論理構造は必ずしもそうではなく，図や脚注を間に挟んで一つの要素を構成する場合がある．物理構造と論理構造の関係をこのように位置づけると，物理構造と論理構造は共通の基本要素をもち，論理要素はひとつ以上の基本要素の並びから構成される．そして基本要素は，複数の論理要素を自分の中に含まないことがその条件となる．

OCR ソフトウェアのレイアウト認識の役割をこのような基本要素を矩形領域として抽出することと捉えた場合，その出力は様々な「誤り」を含む．それらは以下のように分類することができる．

1. 複数の論理構造の要素を含んだ矩形領域が抽出される．例えば，節のタイトルと節の本体，本文と脚注，図や表とそのタイトル，がひとつの矩形領域を構成する．

2. その一部にテキストを含むような図や表を多数の小さなテキスト矩形領域の集まりと認識する．

3. 多段組の文書を前提とすると不必要であるような過分割を行う．箇条書きやタイトルにおいて，中黒等の記号や番号等と本体部分との間隔が広かったり，文章中の句読点の配置等により，矩形の境界と誤認識されるような空白が生じることが原因である．

1. については，スキャンの品質が低く，段組みの間隔が狭い文書などに対しては２段組みの左右の段をひとつの矩形と認識するなど致命的な誤りを犯す場合もあ

[4] http://poppler.freedesktop.org
[5] 著者の調査不足であれば，ぜひご教示いただきたい．

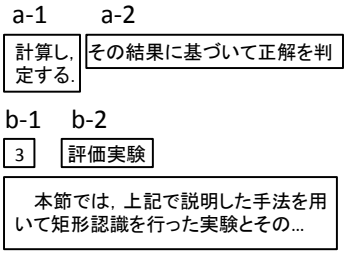

図 3: レイアウト認識の誤り例

る．3. は，図 3 に示すような場合で，a-1 と a-2，b-1 と b-2 は，それぞれひとつの要素とされるのが望ましい．

このような背景に基づき，図 4 に示すような手順で論理構造の抽出を行う．入力は，OCR ソフトウェアの処理結果とする．ディジタル文書の場合は，その文字認識結果にテキスト抽出の結果を重ねあわせて文字認識誤りの訂正を行うことを考えている．

基本要素抽出 OCR ソフトウェアのレイアウト認識の誤り訂正（上述した 3 種類の誤りの訂正）を行い，基本要素を抽出・作成する．

論理種別注釈 得られた基本要素に論理構造の観点からの種別を注釈づける．

論理構造構築 論理種別を注釈づけられた基本要素の並びから論理構造を得る．

4.2 コーパス

これらの処理の仕様検討と評価を目的に，小規模なコーパスを作成した．2003, 2006, 2009, 2013 年からほぼ同数をプログラム構成に基づく種別のバランスのみ考慮して無作為抽出した言語処理学会年次大会予稿 100 件を対象とし，まず，それら文書の e-typist のレイアウト認識の結果を人手により基本要素として適切なものを矩形領域とするように修正した．修正は，前述の「誤り」に対応して以下の 3 つの方針に基づく．

1. 改行で区切られた本文中の式や素性構造表現等については，本文と異なる領域とする，節のタイトルは本文から分離するなど，原則として分割の方向で，基本要素として適切な矩形領域へと修正する．適切な基本要素ということで，これらの矩形には論理種別（後述するように表 1 の `type` 属性の値として示される）のいずれかを付与することができる．

2. 図や表を，図に分類されるひとつの矩形領域とする．それぞれのタイトルは異なる領域とする．

表 1: 論理種別の注釈

属性	値	説明
type	header	ヘッダ
	page	ページ番号
	footer	ページ番号以外のフッタ
	title	論文タイトル
	auth	著者情報（所属等も含む）
	abst	梗概
	stitle	セクション（節）タイトル
	sstitle	サブセクションタイトル
	ssstitle	サブサブセクションタイトル
	body	本文
	list	箇条書き（全体）
	listitem	箇条書き項目
	footnote	脚注
	equ	数式
	fig	図
	tab	表
	figcap	図タイトル
	tabcap	表タイトル
	note	図表註釈
	ack	謝辞（全体）
	acktitle	謝辞タイトル
	ackbody	謝辞本文
	reftitle	参考文献タイトル
	refbody	参考文献本体（全体）
	refitem	参考文献項目
par	whole	全体（デフォルト値）
	first	先頭部分
	mid	中間部分
	last	末尾部分

3. 多段組を前提とした不必要な分割については，可能であれば統合を行う[6]．

その後，矩形領域（=基本要素）に表 1 に示す論理構造に関連するふたつの属性の注釈付を行った．第一の属性 `type` は論理構造における要素の種類（論理要種別）を示すものである．第二の属性 `par` は論理構造の観点ではひとつの要素となるべきものが，物理的制約で分割されているか，分割されている場合は，そのどの部分であるかを示している．

表 1 に示されているように，論理要素の種別においては，箇条書き部分を本文から区別する等，その後の利用で必要と思われるものに対してやや細かい区分がなされている．また，箇条書きや参考文献等において，その項目（listitem, refitem）と全体（list, refbody）の 2 種類の種別を設定している．粒度を揃えるということでは，両方を基本要素とすることは問題であるが，これは自動で行われるレイアウト認識の結果の修正を最小限とするための配慮である．つまり，箇条書きや参考文献の部分をレイアウト認識すると，文書のスペーシングにより，全体がひとつの矩形領域とされる場合

[6] 利用している e-typist では，テキストに分類される領域について，自動認識結果を更に分割することは自由に可能であるが，統合については実行できない場合があり，完璧な修正となっていない場合がある．

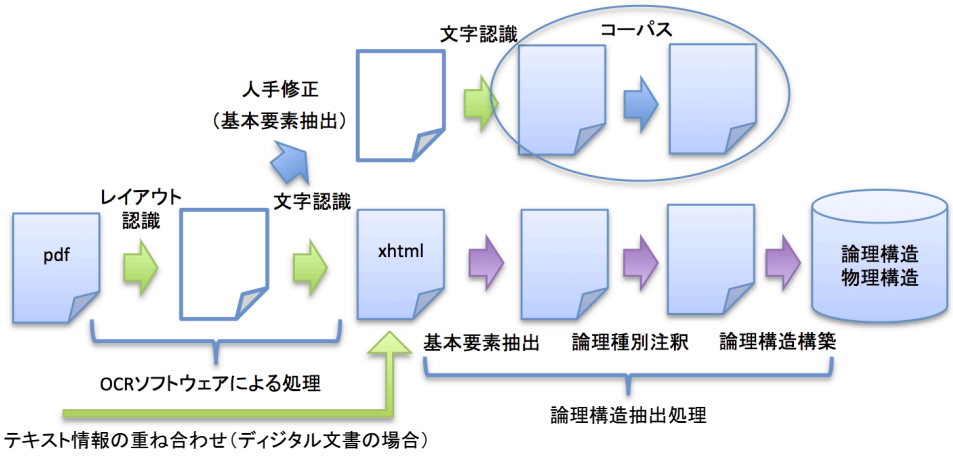

図 4: 論理構造抽出の枠組み

と，項目ごとに矩形領域とされる場合とがある．このいずれの場合も人手修正を行わず，異なる注釈を行うことで対応している．ただし，箇条書き部分が前後の本文と同じ領域とされてしまったり，一部の複数の項目だけがひとつの領域と認識された場合は，領域を分割することで修正を行っている（方針 1.）．

前述の論理構造抽出処理において，基本要素抽出は，レイアウト認識結果修正を模擬することに，論理種別注釈はその後の注釈の模擬に相当する．論理構造構築は，もしそこまでの処理が完璧であれば，単純なパージングであるが，そうでない場合は，処理誤りに起因するノイズへの対応や，場合によっては前段の処理へのフィードバックが必要になる．

4.3 実装

現在，基本要素抽出と論理種別注釈について実装を進めている．

基本要素抽出では，前述の 3 種類の誤りに対し，アルゴリズム的に修正を行っている．1. については，矩形の位置，先頭の文字種（先頭文字が空白であることによる字下げの認識を含む），行末における句点の存在，「謝辞」等のキーワードとの一致，等を用いて分割すべき境界の判定を行う．2. については，矩形の位置や大きさ，フォントの大きさ，矩形領域内の空白の割合等を用いて，テキスト領域ではない矩形を削除する．3. についても，同じ文書の別の部分の認識結果から推定される段組みのパラメータを前提として，不自然な横幅を持つ矩形が判定でき，その周囲にある矩形との位置関係から，統合すべきものが判断できることが多いので，それを用いて統合を行う．

テキストと分類された領域について，その効果を測ると，自動レイアウト認識の結果と人手修正後のコーパスとでは，文書毎のマクロ平均で，精度（修正が必要ない矩形数/自動認識結果での矩形数）が 0.58，再現率（修正されていない矩形数/人手修正後の矩形数）が 0.63 であるのに比較して，自動レイアウト認識結果に基本要素抽出を施したものは，人手修正後のコーパスに対して，精度（両者に共通する矩形数/基本要素抽出後の矩形数）は 0,79，再現率（両者に共通する矩形数/人手修正後の矩形数）は 0.75 と向上する．クローズドテストであり，2013 年のものを主に参照して開発したため，それらについては精度 0.89，再現率 0.90 と高い性能が得られる．一方で，2003 年のスキャン文書については，段組みを誤認識する等，致命的な誤りを含むものも多く，よい結果が得られていない．また図表や式については，複数のテキスト領域と誤って認識されたものから，そこに図表等が存在したことが復元される必要があるが，この処理は現時点では行っていない．

論理種別注釈は，コーパスを用いた機械学習を行い，CRF による系列ラベリングを行っている[7]．矩形領域の位置，先頭の文字種別等とバイグラムの情報を素性としている．10 分割交差検定で．表 2 に示す混同行列が得られている．ここでは，その後の応用を前提とした分類とし，list と listitem，stitle と sstitle 等はまとめている．また，コーパス中の論文には梗概（abst）を含むものが極めて少なかったため表に含めていない．全体の正解率は 87%である．

5 関連研究

PDF 文書からテキストを抽出し，検索を行う試みは幾つか行われている．阿辺川らは，抽出されたテキストと PDF 文書を用いて，参考文献へのリンクやキー

[7]CRF の実装は CRF++（http://taku910.github.io/crfpp/）を用いた．

表 2: 論理種別推定

正解＼推定	ack	at	au	bdy	equ	fig	fc	ft	fn	hd	lst	nt	pg	rb	rt	st	tab	tc	tt
ack	7			3							2								
acktitle (at)		1		1							1				9				
auth (au)			206	1												3			
body (bdy)	1			1860			7		7		157			8		5	1	9	
equ					159	27											10		
fig			1		24	217											36		
figcap (fc)			1	7			240				5							17	1
footer (ft)								107											
footnote (fn)				10					94		13					4			
header (hd)										76									
list (lst)		2		183		2	11		13		676			5	19	32		21	
note (nt)				1			3		1		6	4	1				1	3	
page (pg)							1						301						
refbody (rb)				7					1		19			96	1	1			
reftitle (rt)		1		2							11			1	85	3			
stitle (st)				2		1	1		3		11					1117			
tab					16	33											229		
tabcap (tc)				11			27				9							233	
title (tt)																			96

ワードへの脚注を備えた閲覧システムを実現している [1]．ACL Anthology[8]を対象に，統語解析可能なテキストを得るために，ディジタル文書，スキャン文書の解析が試みられている [3, 13, 14]．得られたテキストを統語意味解析し，意味に基づく検索を実現することがその目的である．増田らは，テキストマニングの対象として，OCR 読み取りを用いたテキストを利用している [10]．数式等を含めたより高精度な復元処理が磯崎によって検討されている [7]．

文書の構造認識については，Klink らや Luong らの研究がある [8, 9]．ここでも CRF を用いて，文書の構成要素からなる論理構造を明らかにしているが，検討されているのは論理種別注釈に相当する部分で，レイアウト認識の誤りに対する処理は含まれていない．文書の構造を利用するという点では前述の阿辺川のシステムに加えて，難波らが引用情報を解析して，その役割を利用した構造化を行っている [11]．

6 おわりに

文書構造に基づく対話的情報アクセスの枠組みを提案し，そのための文書表現を構築するために必要になる文書の論理構造解析について現状を報告した．提案した枠組みはまだ構想段階に留まっており，今後，以下の検討が必要と考えている．

研究論文等に対する検索意図の収集と分析 3 節で考察した検索意図の分類について，現実の検索意図を収集する等を通じて，詳細化を行い，それらの検索意図に応えるための照合方式を検討する．現在想定している文書表現がそのような照合方式に充分であるかを確認する．

閲覧等，インタラクションの枠組み設計 2 節の枠組みにおいて，まだ十分に検討されていない対話的な情報アクセスについて，文書とその部分の行き来や論理構造と物理構造の行き来等，これまでにはない焦点の移動について検討し，基本的な操作を明らかにする．

論理構造の抽出の精度向上と実現 4 節で提案している方式について引き続き検討を進め，どの程度の精度が得られるかの見通しを得る．それを受けて，文書表現の作成にどの程度の人手介入を必要とするか等を考慮に入れて，システム全体の設計を進める．また，現在では異なる方針で実装している基本要素抽出と論理種別注釈について枠組みの融合が可能かを検討する．

いずれも小さくはない課題であるが，順次検討を進めていきたい．

[8] http://aclweb.org/anthology/

参考文献

[1] 阿辺川武, 相澤彰子: 脚注表示機能を備えた論文閲覧システム Sidenoter, 『言語処理学会第 20 回年次大会予稿集』, pp. 796–799 (2014).

[2] Bates, M.J.: The Design of Browsing and Berrypicking Techniques for the Online Search Interface, *Online Review*, Vol. 13, No. 5, pp. 407–424 (1989).

[3] Berg, Ø., Oepen, S., Read, J.: Towards High-Quality Text Stream Extraction from PDF. Technical Background to the ACL 2012 Contributed Task, *Proc. of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pp. 98–103 (2012).

[4] Callan, J.P.: Passage-Level Evidence in Document Retrieval, *SIGIR '94*, pp. 302–310 (1994).

[5] Ellis, D.: A Behavioral Approach to Information Retrieval System Design, *Journal of Documentation*, Vol. 45 No. 3, pp. 171–212 (1989).

[6] Hearst, M.A., Plaunt, C.: Subtopic Structuring for Full-Length Document Access, *SIGIR '93*, pp. 59–68 (1993).

[7] 磯崎秀樹: PDF 中の TEX 記号の復元と ACL Anthology への適用, 『言語処理学会第 19 回年次大会予稿集』, pp. 956–959 (2013).

[8] Klink, S., Dengel, A., Kieninger, T.: Document Structure Analysis Based on Layout and Textual Features, *Proc. of International Workshop on Document Analysis Systems, DAS2000*, pp. 99–111 (2000).

[9] Luong, M., Nguyen, T., Kan, M.: Logical Structure Recovery in Scholarly Articles with Rich Document Features, *International Journal of Digital Library Systems*, Vol. 1, No. 4, pp. 1–23 (2010).

[10] 増田勝也, 丹治信, 植松すみれ, 美馬秀樹: 研究動向分析のための論文のデジタルテキスト化とマイニングシステム, 『言語処理学会第 20 回年次大会予稿集』, pp. 792–795 (2014).

[11] 難波英嗣, 神門典子, 奥村学: 論文間の参照情報を考慮した関連論文の組織化, 『情報処理学会論文誌』, Vol. 42, No. 11, pp. 2640–2649 (2001).

[12] Salton, G., Allan, J., Buckley, C.: Approaches to Passage Retrieval in Full Text Information Systems, *SIGIR '93*, pp. 49–58 (1993).

[13] Schäfer, U., Read, J., Oepen, J.: Towards an ACL Anthology Corpus with Logical Document Structure. An Overview of the ACL 2012 Contributed Task, *Proc. of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pp. 88–97 (2012).

[14] Schäfer, U., Weitz, B.: Combining OCR Outputs for Logical Document Structure Markup. Technical Background to the ACL 2012 Contributed Task, *Proc. of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pp. 104–109 (2012).

[15] Tellex, S., Katz, B., Lin, J., Fernandes, A., Marton, G.: Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering, *SIGIR '03*, pp. 41–47 (2003).

[16] 上田修一, 倉田敬子: 『図書館情報学』, 勁草書房, pp. 217–218 (2013).

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第10回)
SIG-AM-10-02

動向に関する問いに答える コンテクスト検索エンジンの開発

Development of Context Search Engine Focusing on Trend-related Queries

高間　康史[1]　Yanjun Zhu[1]　桑折　章吾[1]　山口　晃一[1]　瀧口 慈勇[1]

Yasufumi Takama[1], Yanjun Zhu[1], Shogo Kori[1], Koichi Yamaguchi[1], Satoru Takiguchi[1]

[1] 首都大学東京大学院システムデザイン研究科
[1]Graduate School of System Design, Tokyo Metropolitan University

Abstract: This paper introduces a context search engine designed for answering trend-related queries. Aiming at narrowing the gap between user's information need and functions provided by an existing search engine, we are developing advanced search engine that focuses on the task of answering trend-related queries. As the task of answering trend-related queries is supposed to be common in various domains, we expect it could be used for various purposes. After explaining the structure and function of the proposed search engine, its potential application and the possibility of improvement are discussed.

1. はじめに

本稿では，動向に関する問いを対象としたコンテクスト検索エンジンについて概説し，想定する活用方法や今後の開発方針について考察する．

Web 上に存在する多種多様なリソースへのアクセス手段として，検索エンジンが現在広く用いられている．検索エンジンが一般的な存在となった理由として，「指定したキーワードを含む Web ページを見つける」という基本検索機能が，直感的で検索スキルのないユーザにとってもわかりやすいことが挙げられる．また，この基本検索機能がドメイン・タスクによらず広く一般的に利用可能であること，複数の検索（クエリ）を組み合わせることで，多様な用途に利用可能であることなども検索エンジンの利点といえる．

しかしその反面，検索エンジンが提供する基本検索機能は低レベルにとどまっており，ユーザの抱く検索要求との乖離が大きくなっていると考える．すなわち，多種多様な情報要求を，検索エンジンに入力すべき一連のクエリ（キーワード）に分割する必要があり，一般ユーザにとっては簡単な作業でない[1,2]．また，熟練者にとっても効率的な情報アクセスを阻む要因となっていると考える．

この問題に対し，動向に関する問いにタスクを限定することで，現在の検索エンジンよりも高度な検索機能を提供するコンテクスト検索エンジンを開発している[3,4,5]．動向に関する問いは幅広いドメインにみられるものであるため，既存検索エンジンと同様ドメインによらず利用可能であることが期待できる．例えば，最新のニュースに気になる話題があった場合に，過去に同様の話題が注目を集めたことがあったか調べるといった，気軽な用途も考えられる．また，データセットの組合せが価値を創出するデータ市場[6]において，多様なリソース間の潜在的関係を見いだすツールとしても利用可能と考える[7,8]．

本発表では開発中のコンテクスト検索エンジンについて紹介するとともに，想定する活用方法，および今後の開発における課題について述べる．

2. 関連研究

2.1. サーチエンジンの高度化

既存検索エンジンの知的化・高機能化を目指す研究はこれまでにも様々に試みられている．代表的なアプローチとしては，可視化によるインタフェースの改良[9,10]，自然言語によるクエリ入力を受け付けるアプローチ[11,12]，検索対象とするドメインを限定し，専門検索エンジンを構築するアプローチ[13,14]などが研究されている．

情報可視化を利用したアプローチでは，クエリ入力を支援する GUI[10]や，検索結果をクラスタリングして提示するといったインタフェース[9]の改良が研究されてきた．クラスタリングを利用した検索

エンジンは，Vivisimo や Grokker，Kartoo などが公開されていたが，定着せずに現在に至っている．

自然言語によるクエリ入力は，キーワードではなく文として情報要求を表現可能であるというだけでなく，検索結果として直接的な回答を期待することが暗黙に含まれていると言える．従って，自身の情報要求を複数のクエリに分解することで必要な情報を得る，既存検索エンジンとは異なるアプローチである．直接回答を得るアプローチも利用価値の高いものと言えるが，利用者の創意工夫により多様な情報を得ることのできる，現在の検索エンジンと同様のアプローチも大事であり，継承していくべきと考える．

専門検索エンジンに関する研究として，亀井らは，Web 上に存在するソフトウェア開発に関する知見や情報を対象とした検索エンジンを提案している[13]．Web 上に存在するソースコードや付属するドキュメント，Tips などのソフトウェア資源をクローリングにより収集し検索可能としている．ソースコードを解析し，索引付けすることで，クラス名，引数や返値の型，行数などを指定した検索を可能としている．

小久保らは，「検索隠し味」と呼ぶドメインを限定した専門検索エンジンの構築手法を提案している[14]．決定木学習を用いて Web ページ集合から抽出したブール式を，ユーザが入力したクエリに加える事で，既存検索エンジンの検索結果を特定ドメインに特化させている．

これらの検索エンジンは，検索対象ドメインをある領域に特化させることで，既存検索エンジンよりも効率的な検索の実現を目指している．これに対し，本稿で紹介するコンテクスト検索エンジンでは，「動向に関する問い」という，ドメインに依存しないタスクを対象とすることで，広く一般に利用可能という既存検索エンジンの特徴を継承するとともに，対象タスクに特化した高機能な基本検索機能の実現を図る点で異なる．

2.2. 動向情報

動向情報とは，ある商品の価格や売上の状況，ある会社の業績状況などの時系列データを基として，その変化を通時的にとらえつつ，総合的にまとめ上げることで得られるものであり[15]，様々なタスク・ドメインにおいて意思決定の材料として用いられている．近年，LOD（Linked Open Data）[16]などとして公開されるデータの中にも動向情報は多数存在し，その活用が期待されている．田代らは，時間に関連する属性を持つリソースを抽出し，ヒストグラムを描画するツールを提案している[17,18]．

松下らは，動向情報が含まれるテキストを視覚情報として要約することを目的として，テキストに含まれる情報を用いてグラフを描画する方法を提案している[19]．石黒らは，異種情報間の時間的関連性についての検索をコンテクスト検索と定義し，コンテクスト検索に基づく対話的な時系列データ分析を支援するシステムを提案している[20]．為替レートデータとニュース記事の見出しを対象データとして類似変動区間検索機能，類似イベント検索機能を基本検索機能として提供している．加藤らは，検索数やヒット数など，Web 上の動向に関連する基本情報を Web コンテクスト情報として定義し，これらに基づく同時期流行アイテムの検索手法を提案している[21]．

3. コンテクスト検索エンジン

3.1. システム構成

図 1 に，開発中のコンテクスト検索エンジンの構成を示す．実装には Ruby on Rails3.2，Apache2.2，MySQL5.0 を用いている．クローラー（Crawler）は Web 上で公開されている動向情報を収集し，検索対象とする特徴的な動向変動を計算し，データベース（DB）に格納する．Web サーバ（Web Server）はクライアント（Client）からのクエリを受け付けてデータベースを検索し，検索結果をクライアントへ返す．クライアントとしては通常の Web ブラウザからのアクセスを想定する他，任意アプリケーションからの利用も可能となるように API も実装している．

3.2. 動向データの収集

開発中のコンテクスト検索エンジンでは，Web から収集可能な動向情報を以下の二種類に大別し，収集している．

- Web コンテンツとしての動向データ：各アイテムの価格や販売量に関する統計データの様な，各企業や組織・団体によりコンテンツとして公開される動向情報
- Web 利用としての動向データ：各アイテムをキーワードとして既存検索エンジンで検索した際のヒット数や，ブログ記事数などといった，Web 上でのユーザ活動により発生する動向情報

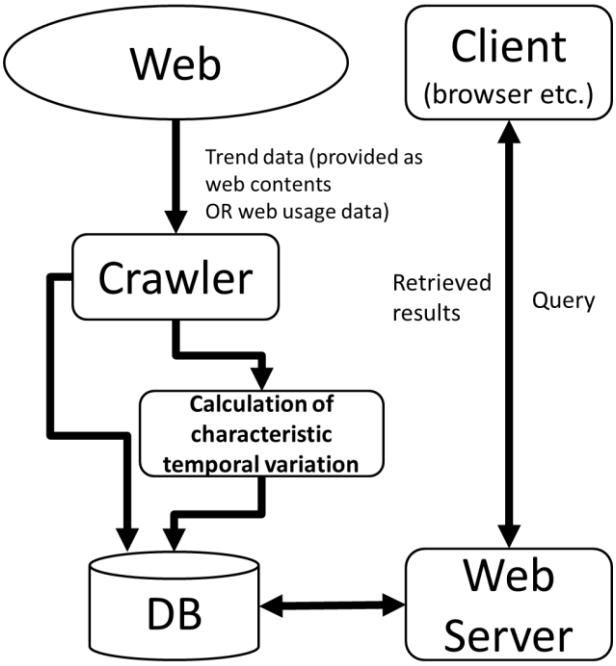

図 1. コンテクスト検索エンジンの構成図

開発中のコンテクスト検索エンジンでは，前者として総務省統計局から人口や雇用者に関する統計データなどを収集している．また，後者として Google Trends の検索数などを収集している．現在検索可能な動向データ数を表 1 にまとめる．なお，Web コンテンツ，Web 利用データ両方のリソースを持つアイテムが存在するため，アイテム数の合計は両データのアイテム数の和よりも小さくなっている．

表 1. 収集した動向データの概要

	Web コンテンツ	Web 利用	合計
アイテム数	179	27,690	27,848
リソース数	186	28,426	28,612

3.3. 検索機能

コンテクスト検索エンジンでは,「既存検索エンジンよりも動向に関して高度な検索が可能であること」,「複数の検索を組み合わせることで，動向に関するユーザの多様な問いに答えられること」を設計方針としている．これらを満たすために，以下の 3 種類の基本検索機能を実装している．

(1) 指定したアイテムに関する動向（リソース）が特徴的変動を示した期間の検索
(2) 指定した期間に特徴的変動を示したアイテム・動向の検索
(3) 指定したアイテムに関する動向が特徴的変動を示した期間に同様の変動を示したアイテム・動向の検索

変動に関しては，現状では以下の 6 種類について利用可能であるが, 今後追加をしていく予定である．

- 最大値（MAX）／最小値（MIN）: 各動向情報が最大値／最小値を取る月
- 急上昇（SI）／急下降（SD）: 3 ヶ月以内に，その動向情報の|最大値−最小値|の 1/5 以上の単調増加／減少が見られる期間
- 山形（PEAK）／谷形（BOTTOM）: その動向情報の|最大値−最小値|の 1/10 以上の単調増加／減少が見られた後，減少／増加に転じた期間

クエリの例を以下に示す．

・[自転車 PEAK @period] : 自転車（アイテム）に関する何らかの動向が山形となった期間の検索
・[2008/05-12 BOTTOM @item] : 2008 年 5〜12 月の間に何らかの動向が谷形となったアイテムの検索
・[iPad S+ヒット数 MAX @item] : iPad のヒット数が最大となる期間に同じ変動をしたアイテムの検索

最後の例で,「S+ヒット数」は検索対象とする動向を指定している．

クエリの入力に関して，初期のコンテクスト検索エンジンでは上記クエリをユーザが直接入力する形式を採用していた．それでも正しいクエリが入力される割合は商用検索サービスと同程度であることを確認しているが[5]，フォーム形式を採用したインタフェースも開発している[22]．フォーム形式を採用したインタフェースのスクリーンショットを図 2 に示す．変動タイプおよび出力タイプについてはプルダウンメニューから選択して指定可能となっている．

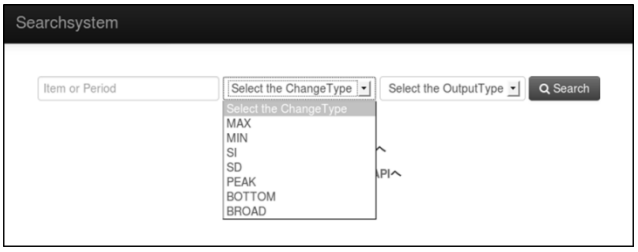

図 2. フォーム形式のクエリ入力インタフェース

図 3 に，検索結果画面のスクリーンショットを示す．現状ではランキング機能はなく，クエリを満た

すアイテムや期間などが列挙される．各検索結果はアイテム名，リソース名，クエリを満たす期間，当該情報が都道府県などに関するものの場合は該当地域から構成される．アイテム名をクリックすることで，その動向の折れ線グラフが表示される．また，各検索結果の右端には Google 検索へのリンクがあり，これを利用してアイテム名＋動向名をクエリとし，期間をオプションとして指定した Web 検索を行うことができる．これにより，該当時期の Web 上での話題などを調べることが可能である．

図 3. 検索結果画面のスクリーンショット

4. コンテクスト検索エンジンの活用と機能拡張

4.1. データリソース間の関係発見

コンテクスト検索エンジンの活用例の一つとして，異なるデータリソース間の関係発見に利用することを検討している．官公庁や地方公共団体を中心とするオープンデータの流れや，ビッグデータなどのキーワードに代表されるデータ活用への意識の高まりにより，異なるデータを組み合わせて新たな価値を創造する必要性が指摘されており，データ市場に対する関心が高まっている．データ市場においてやりとりされるデータリソース（データセット）の中には内容を公開できないものも存在するため，内容を公開することなく，その価値を見積もることを可能とするためにデータジャケットの概念が提案されている[23]．データジャケットはデータリソースの変数名といったメタデータや概要を記述したものであり，これを利用することで価値を生み出すデータリソースの組合せなどを検討する．IMDJ (Innovators Marketplace on Data Jackets)ではデータジャケットを利用し，市場の多様な利害関係者がワークショップ形式で議論を通じながら自身の問題解決に繋がるデータリソースの組合せを発見する．一般に，データリソース間の関係を見つけるためには，共通あるいは関連するインスタンスに着目したり，関連する属性に着目するなどのアプローチが一般的と考えられる[8]．これに対し，コンテクスト検索エンジンを利用した場合には，動向情報の関連性の観点からデータリソース間の関係を発見することが期待できる．同時期に流行したなどの時間的関連性は，時系列性のあるリソースで，データ収集期間にオーバラップがあれば計算可能であるため，より多様なデータリソース間の関係発見に貢献することが期待できる．

これまで，開発者および実験協力者がコンテクスト検索エンジンを利用し，動向情報の観点からアイテム間の関係を発見することを試みている．これまでに発見した事例をいくつか紹介する．前掲の図 3 は，インフルエンザと同時期に動向情報が急上昇するアイテムの検索結果である．ここで，急上昇する期間は複数存在することがあり得るため，検索結果には同じアイテム・リソースが複数回出現している．図より，インフルエンザと同時期に動向情報（検索件数）が急上昇するアイテムとして，空気清浄機が検索されていることがわかる．これは，空気清浄機の高機能なものには，インフルエンザへの効果をうたったものがあることに対応している．

この他，以下のような関連アイテムが発見されている．

(1) 原発と自転車
(2) カメラとビデオカメラ
(3) キャベツとトマト
(4) いちごとフグ
(5) 炊飯器と JR 西日本

(1) に示した二つのアイテムは，共に 2011 年 3 月から 12 月の間に動向情報が最大値を迎えている．当該期間は東日本大震災直後であり，原発の検索数が検索結果に含まれているのは妥当な結果と言える．一方，自転車は販売量に関する動向情報が当該期間に最大値を迎えていた．当時のニュース記事などを確認したところ，交通機関が止まった場合の交通手段や，省エネのために自転車を購入する人が増加しており，それが反映した結果と言える．原発と自転車の間には一見関係はないように考えられるが，動向を切り口とすることで，自転車販売量と原発検索

数という異なるデータリソース間の関係が発見できた事例といえる．

上記の例は，あるイベント（東日本大震災）が共通の原因となって，同時期に同様の動向変動が見られたものである．同様の根拠により関係が発見された事例として，(2) に示す二つのアイテムは，2012年 4〜5 月に価格が高騰していた．この原因としては，2011 年にタイで発生した洪水により，電子機器の部品工場が多数被害にあったことが考えられる．カメラとビデオカメラは元々関連の深いものと言えるが，同様の特徴的な変動が観測された原因としては興味深いものと考える．(3) のキャベツとトマトの例では，天候不順のため同じ時期に価格が高騰していることによって関連性が生まれており，同様の根拠に基づくものと言える．

上記とは異なる根拠に基づく関連性として，(4) の例では周期性のある動向変動が根拠となって関係が発見されている．例えば，いちごとフグは旬や収穫時期が 3, 4 月と一致しており，その時期に価格が下落していることにより動向情報上の関連が生まれている．

二つのアイテムに直接関係する話題が発生したことによって関連性が生まれるケースも見られた．(5) に示した炊飯器と JR 西日本に関しては，「JR 西日本商事が今春で引退する特急電車を模した炊飯器を発売」というジョーク画像がネット上で話題となり，両アイテムの検索数が上昇したことが原因となっている．

この様に，一口に動向情報と言っても，多様な根拠に基づく関係の発見が可能であり，異なるデータリソース間の関係に気づくきっかけとして活用できると考えている．

4.2. 機能拡張に向けての考察

コンテクスト検索エンジンの設計方針は，「幅広いドメインに適用可能であり，利用者の創意工夫により多様な情報要求を満たすことができる」という現在の Web 検索エンジンの利点を継承しつつ，タスクを動向に関する問いに答えることに限定することで，より高度な基本検索機能を提供することである．これを踏まえ，今後の機能拡張などについては以下に取り組む必要があると考えている．

(1) 検索エンジンとしての機能拡張
(2) データベースの拡充
(3) 活用方法の検討

検索エンジンのとしての機能拡張に関しては，変動タイプの追加といった，コンテクスト検索エンジンに特有の機能拡張を検討している．この他，既存の Web 検索エンジンとのアナロジーにより，実装すべき機能について検討することで，既存検索エンジンの良さを継承可能と考えている．例えば，現在の検索エンジンでは，検索結果はランキングされてユーザに提示される．これにより，ユーザは欲しい情報を効率よく発見できている．また，ランキングは検索エンジンをデータベース検索と区別する大きな特徴でもあると考える．データベース検索では，利用者が検索したいものが満たす条件を具体的に指定する．また，検索結果をソートする場合もその条件は利用者が指定する．これに対し検索エンジンでは，検索オプションとして AND, OR などを指定したり，ファイルタイプやドメインなどを限定することもできるが，データベース検索ほど詳細なものではない．また，ランキングに関しては利用者が条件を指定する必要はない．すなわち，事前の検索意図はある程度漠然としていて，検索結果を見て発見するという行為が前提となっているのが検索エンジンであると言える．従って，開発中のコンテクスト検索エンジンも，ランキング機能を導入することが必須と考えている．

現在の検索エンジンでは，多様な要因を考慮してランキングが決定されていると言われている[24]．また，これらの多様な要因は，ランキング学習により統合され，スコアを決定する関数が決定される[25]．コンテクスト検索エンジンにおいても，時系列データとしてみた場合の特徴や，クエリとの適合性など多様な要因について検討し，ランキングを導入することを計画している．

検索エンジンに近年導入された拡張としては，スニペット[26]，クエリ推薦が挙げられる．スニペットは Web ページ中でクエリに指定された単語を含む部分を抽出し，検索結果の一部として提示されたものである．スニペットにより，指定した単語が Web ページ中でどのように出現するかがわかるため，検索結果画面から実際の Web ページへ遷移することなしに結果を吟味することが可能となる．このことは効率的な情報発見に貢献している．コンテクスト検索エンジンにおいては，現在は別画面として提示している動向情報の折れ線グラフをスパークラインとして検索結果画面に描画することで，スニペットの役割を果たすことが期待できるため，現在実装を進めている[22]．

クエリ推薦は，クエリに追加することで検索結果の絞り込みに有効であることが期待できるキーワードを利用者に提案する技術であり，クエリログを利用して生成される．すなわち，検索におけるベスト

プラクティスの共有と見ることもできる．コンテクスト検索エンジンにおいては，複数の基本検索機能を提供し，これらを組み合わせて多様な情報要求を満たすことを想定している．その様な検索の組み合わせを誘発するためには，現在入力中のクエリに対する推薦だけでなく，次に実行すると良いクエリを提案することも重要と考え，現在その推薦手法を検討している．

(2)に挙げたデータベースの拡充に関しては，検索可能なアイテム数やリソース数の増加が挙げられる．検索可能なアイテム数を増加させるためには，多数のアイテムに関する動向情報を含む巨大なリソースを取り入れることが効果的であり，Wikipedia のページビューデータ[27]を検索可能にする準備を現在進めている．リソース数の増加は，4.1 節に示したデータリソース間の関係発見においても，意外な関連性を見つけるうえで重要と考えている．この時，異なる Web サイトでは，それぞれ異なる様式でデータが公開されていることが一般的であるため，ラッパー構築のコストが問題となる．従って，SPARQL で統一的にアクセス可能な LOD はラッパー構築コストの観点から魅力的であり，導入を検討したいと考えている．

(3)に挙げた活用方法に関しては，現在は 4.1 節に挙げた関係発見を中心に考えているが，気軽かつアドホックな利用も含め，多様な活用方法について検討をしていきたいと考えている．そのためには，コンテクスト検索エンジンを継続的に運用し，利用事例を収集することが効果的であるため，公開に向けた整備を進めている．

5. おわりに

本稿では，動向に関する問いに答えることに特化したコンテクスト検索エンジンについて概説し，その活用や今後の機能拡張の方向性について考察した．コンテクスト検索エンジンは，幅広いドメインに適用可能という既存検索エンジンの特徴を継承しつつ，タスクを動向に関する問いに答えることに限定することで，より高度な基本検索機能を提供することを目的としている．利用者の創意工夫を引き出し，多様な情報要求を満たすことを支援できるような検索エンジンの実現を目指し，本稿で考察したような機能拡張に取り組んでいく予定である．

謝辞

本研究の一部は JSPS 科研費 24650040，15H02780 の助成による。

参考文献

[1] A. Spink, D. Wolfram, M. B. J. Jansen, T. Saracevic, Searching the Web: The Public and Their Queries, Journal of the American Society for Information Science and Technology, Vol. 52, Issue 3, pp. 226-234, 2001.

[2] 齋藤，三輪，Web 情報検索におけるリフレクションの支援，人工知能学会論文誌, Vol. 19, No. 4, pp. 214-224, 2004.

[3] 加藤，桑折，高間，「動向に関する問い」を対象タスクとしたコンテクスト検索の提案, 人工知能学会第 3 回インタラクティブ情報アクセスと可視化マイニング研究会，pp.7-12，2013.

[4] 桑折，加藤，高間，検索エンジンを用いた情報検索におけるユーザ行動の分析，人工知能学会第 4 回インタラクティブ情報アクセスと可視化マイニング研究会，pp.9-14，2013.

[5] 高間，加藤，桑折，石川，動向に関する問いを対象とした検索エンジンの提案，人工知能学会論文誌，Vol. 30, No. 1, pp. 138-147, 2015.

[6] C. Liu, Y. Ohsawa, Y. Suda, Valuation of Data through Use Scenarios in Innovators' Marketplace on Data Jackets, IEEE 13th International Conference on Data Mining Workshops (ICDMW), pp. 694-701, 2013.

[7] Y. Zhu, Y. Takama, Y. Kato, S. Kori, H. Ishikawa, Introduction of Search Engine Focusing on Trend-related Queries to Market of Data, MoDAT2014 in ICDM2014, pp. 512-516, 2014.

[8] 高間，諸，桑折，山口，動向に関する問いに答えるコンテクスト検索エンジンのデータ市場への応用に関する検討，人工知能と知識処理研究会，AI2014-26, pp. 5-8, 2014.

[9] E. D. Giacomo, W. Didimo, L. Grilli, G. Liotta, Graph Visualization Techniques for Web Clustering Engines, IEEE Trans. Visualization and Computer Graphics, Vol. 13, No. 2, pp. 294-304, 2007.

[10] S. Jones, VQuery: a Graphical User Interface for Boolean Query Specification and Dyamic Result Preview, Working Paper 98/3, Department of Computer Science, University of Waikato, New Zealand, 1998.

[11] A. Ferreira, J. Atkinson, Intelligent Search Agents Using Web-Driven Natural-Language Explanatory Dialogs，IEEE Computer, Vol. 38, No. 10, pp. 44-52, 2005.

[12] 徳永，言語処理を利用した知的情報アクセス―検索，抽出，要約，分類，QA，オペレーションズ・リサーチ 経営の科学，52(11)，pp.713-718，2007.

[13] 亀井，門田，松本，WWW を対象としたソフトウェア検索エンジンの構築，電子情報通信学会技術

研究報告ソフトウェアサイエンス, Vol. 102, No. 617, pp. 59-64，2007.

[１４] 小久保，小山，山田，北村，石田，検索隠し味を用いた専門検索エンジンの構築，情報処理学会論文誌，Vo l. 43，No. 6，pp. 1804-1813，2002.

[１５] 加藤，松下，平尾，動向情報の要約と可視化に関するワークショップの提案，情報処理学会研究報告／自然言語処理研究会報告, 2004(108), pp. 88‐94, 2004.

[１６] オープンデータと Linked Open Data，情報処理，Vo l. 54, No. 12, pp. 1204-1210, 2013.

[１７] 田代，高間，RDF データベースを対象としたデータ分析支援ツールの提案，第 5 回情報アクセスと可視化マイニング研究会, SIG-AM-05-02, 2013.

[１８] Y. Takama, K. Tashiro, Proposal of Support Tools for Analyzing RDF Database Using TETDM, SCIS&ISIS2014, pp. 1494-1499, 2014.

[１９] 松下，加藤，数値情報の補填とグラフ概形の示唆による複数文書からの統計グラフ生成，知能と情報，Vol. 18, No. 5，pp. 721‐734，2006.

[２０] Y. Takama, K. Ishiguro, Support of Exploratory Analysis of Exchange Rate Data Based on Context Search and Granularity-dependent Similarity Calculation of Temporal Data, International Journal of Affective Engineering, Vol. 13, No. 4, pp. 235-244, 2014.

[２１] 加藤，高間，Web コンテクスト情報に基づく同時期流行アイテム検索手法の提案, FSS2012, pp. 115‐118, 2012.

[２２] 山口，諸，桑折，高間，コンテクスト検索エンジンのインタフェース向上に関する検討，JSAI2015, 1I3-OS-10b-1, 2015.

[２３] Y. Ohsawa, H. Kido, T. Hayashi, C. Liu, Data Jackets for Synthesizing Values in the Market of Data, Procedia Computer Science, Vol. 22, pp. 709-716, 2013.

[２４] M. Tober, L. Hennig, D. Furch, SEO Ranking Factors and Rank Correlations 2014 - Google U.S.-, searchmetrics Whitepaper, 2015.

[２５] 数原，片岡，素性推定器を用いたランキング学習，JSAI2010, 2A1-04, 2010.

[２６] E. Cutrell, Z. Guan, An eye-tracking study of information usage in Web search: Variations in target position and contextual snippet length, CHI’07, pp. 407-416, 2007.

[２７] 吉田，荒瀬，角田，山本，検索頻度推定のための Wikipedia ページビューデータの分析，JSAI2015, 2I1-1, 2015.

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第10回)
SIG-AM-10-03

SOM を利用した Exploratory Search のためのユーザインタフェース開発

Development of the user interface for Exploratory Search using the SOM

徳永 秀和[1]　井上 雄翔[1]

Tokunaga Hidekazu[1] and Inoue Yusho[1]

[1] 香川高等専門学校

[1] National Institute of Technology, Kgawa College

The important thing in Exploratory Search is that a retrieving person clarifies the goal of search. For that purpose, first it is required to find the keyword which related to Search-word. Then, a retrieving person finds the related keyword that he is interested in. However,since the information acquired by search is huge, it is difficult to find the keyword which fulfills conditions from the information. Then, I thought that such a problem was solvable by developing the tool which extracts only required information from search results and displays the clustered result. In order to make a clustering result intelligible visually, a selforganization map is used, and information is arranged and displayed on a two-dimensional map. Moreover, in order to be able to reflect a user's idea in a clustering result, it enables it to change freely the parameter of the feature vector used by SOM. Finally, evaluating the usefulness of this tool by experiment.

１．はじめに

近年の高度情報化にともなってインターネット上の Web ページは急激に増加しており，現在は 1 兆ページを超えるといわれている[1]。この膨大な Web ページの中から必要な情報を得るために，検索の手法は多様化している。なかでも注目されている検索手法が Exploratory Search である。

Exploratory Search とは，情報のニーズが明確でない検索者が，検索で得られる情報を基に検索の目標を明確化しながら，新しい知識を獲得していく検索手法である[2]。検索の目標を明確化するときに重要となるのが，検索語と関連するキーワードである。検索で得られた情報の中から検索者が興味のあるキーワードを見つけ，そのキーワードを基に検索を繰り返すことが目標の明確化につながる。

インターネット検索を行う際の Web ページ滞在の調査によると，検索者が１ページに滞在する平均時間は約１分といわれている[3]。１ページあたりにかかる閲覧時間はそう長くないが，情報ニーズがあいまいで，検索キーワードに対する予備知識の少ない検索者が１ページずつ情報を探索していくと，検索に長い時間を要してしまう。さらに前述したように Web ページの数は膨大であるため，多くの情報の中から検索者にとって本当に有用なキーワードや Web ページを見つけるのは困難であると予想される。したがって，検索情報の中から必要な情報を抽出し，分類して検索者に提示するツールが必要であると考えられる。

そこで本研究では，Web ページから必要な情報を抽出して，それらをクラスタリングして表示することで，Exploratory Search の支援を行う GUI システムを開発することを目標とした。

２．目標達成の手段

Exploratory Search において検索目標を明確化するとき重要となるのが，検索キーワードに関連し，検索者の興味を引くキーワードを見つけることである。本システムでは検索者にそのようなキーワードを見つけやすくすることで，Exploratory Search を支援する。

検索者が特定のキーワードを見つけるためには，まず Web ページ内の情報を絞り込むことが必要であると考えられる。そこで本システムでは Web ページ中の名詞に注目し，それらを検索者の興味を引くキーワードの候補として抽出して，クラスタリングする。また，検索者によって興味を引くキーワードは異なるため，システムが独自に設定するパラメー

タによるクラスタリングの結果が必ずしも興味を引くキーワードの特定につながるとは限らない。そこで自己組織化マップと GUI を組み合わせ，検索者が独自の判断でパラメータを変更してクラスタリングを行うことでキーワードを絞り込むことのできるシステムを開発する[4]。

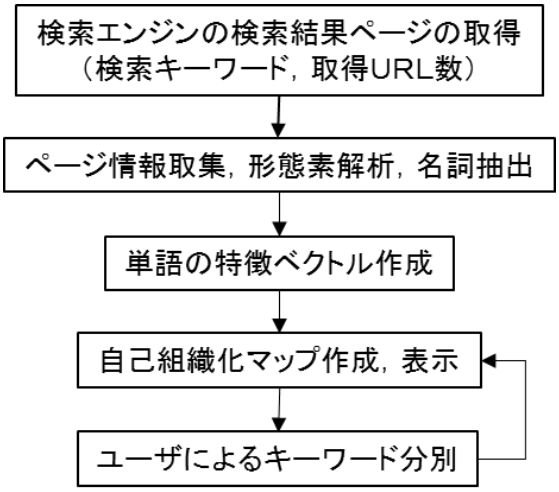

図１　システムの処理の流れ

３．システム構成

3.1 処理の流れ

システムの処理の流れを図１に示す。①キーワードと取得するホームページ数を指定し，検索エンジンより検索結果の HTML 文書を取得する。②HTML 文書より必要なテキスト情報を抽出し，形態素解析し，必要な名詞情報を抽出する。③抽出した名詞（キーワード）の特徴ベクトルを作成する。④キーワードの特徴ベクトルより自己組織化マップを作成し，表示する。⑤ユーザが自己組織化マップのノードを操作し，キーワードを選別する。⑥選別情報を基に，再び自己組織化マップを作成，表示する。⑦これ以降，⑤，⑥を繰り返し，興味を持つキーワードを探索する。

3.2 クラス構成

クラスの構成を図２に示す．SOMtest クラスで全体の流れを制御する。MakePagedata クラスにより，検索エンジンからの HTML 文書取得と名詞データの管理を行う。HTML 文書取得には Httpclient.jar，HTML 文書の処理には jericho-html.jar を使用する。形態素解析は jgo.jar を使用する。自己組織化マップの処理は，ExecSOM クラスが JRI.jar を使用し R の som ライブラリを利用する。SOMgui クラスによりキーワード選別と再自己組織化マップ作成を行う。

3.3 検索結果の取得

システムを開始すると検索キーワード，検索ページ数，また Web ページ本文とスニペットのどちらを

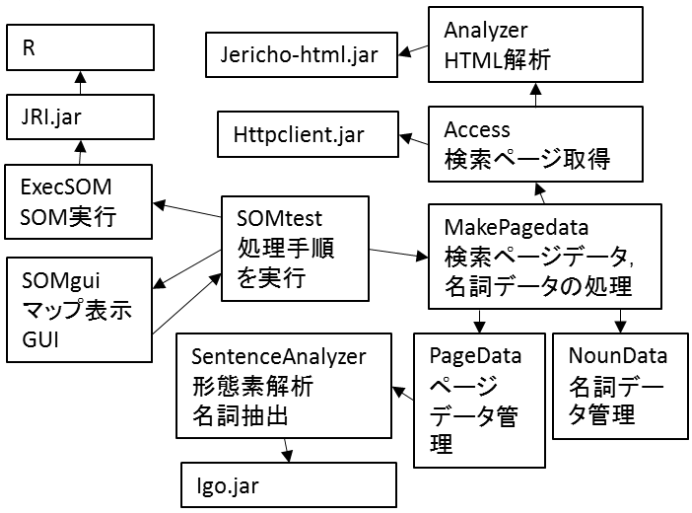

図２　クラス構成

使用するかを入力する画面が表示される。それぞれのデータを入力して実行ボタンをクリックすると，Google 検索エンジンから検索結果の HTML 文書を取得する。本システムでは Google 検索エンジンから検索結果を取得する際に使用する，HTTP ユーザーエージェントというパラメータを固定している。これにより，システム実行環境に依存せず検索結果を得ることができる。

3.4 形態素解析と名詞抽出

検索エンジンから得た検索結果を形態素解析し，形態素の中から名詞のみを，検索キーワードとの関連キーワードとして抽出する。形態素の品詞は階層構造で分類されており，単に名詞といっても数十種類に細かく分類される。本システムでは名詞の中でも特に単独で強い意味を持つことの多い「名詞，一般」と「名詞，固有名詞」を主として抽出する。また「ノンアルコール」などのように，単語として意味を成すが，「ノン」と「アルコール」という複数の形態素に分解されるような単語については，「ノンアルコール」というように一つの単語を関連キーワードとして抽出する。「環境汚染問題」のように複数の名詞が連続する複合名詞は，複合名詞をキーワードとする。

3.5 特徴ベクトル

特徴ベクトルとは関連キーワードの特徴を数値化して並べた多次元ベクトルのことである。本システムでは抽出した全ての関連キーワードについての特徴ベクトルを作成する。特徴ベクトルの属性は，「検索結果全体での出現回数」，「固有名詞であるか否か」，「キーワードの文字数」，「Web ページ１での出現回数」，・・・「Web ページ n での出現回数」である。

3.6 自己組織化マップ

自己組織化マップとは，与えられた特徴ベクトルからそれぞれのキーワードの類似度をマップ上の距離で表現するものである。自己組織化マップ上では類似度の高いキーワードどうしは近くに，類似度の低いキーワードどうしは遠くに配置される。多次元データを持った関連キーワードを 2 次元マップ上に視覚的にわかりやすく表示できるため，多数の関連キーワードを分類し表示する必要のある本システムに適していると考えられる[5]。

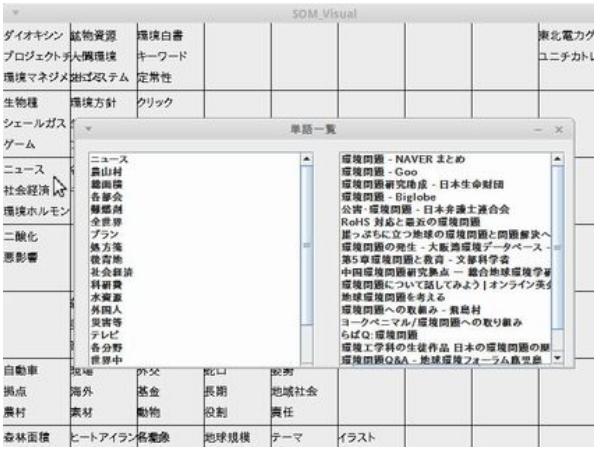

図３　自己組織化マップ

自己組織化マップは統計解析ソフト R によって作成する。本システムで作成・表示される自己組織化マップは，クラスタ数が 10×10，クラスタ形状が四角形のものである。自己組織化マップの作成と同時に，各関連キーワードの重要度の計算が行われる。関連キーワードの重要度は以下の式によって計算される。

重要度 ＝ A＊出現回数＋B＊文字数 ＋ C＊固有名詞
（固有名詞は，固有名詞なら１，違えば 0）

ここで A,B,C,は関連キーワードの各属性の係数であり，ユーザーが独自に設定できる値である。

自己組織化マップの作成と関連キーワード重要度の計算が終わると，図３の画面（単語一覧のポップアップは除く）が表示される。各ノードのマスごとで関連キーワード重要度の高いキーワードが最大３つまで表示される。また各ノードのマスを左クリックすると，図３（単語一覧のポップアップ）ようにクリックしたノード内の全ての関連キーワードと，それらのキーワードを含む Web ページのタイトル一覧を表示した画面が現れる。画面内左側にリスト表示された関連キーワードをクリックで選択すると，選択した関連キーワードが含まれる Web ページのみが右のリストに表示される。このとき関連キーワードは複数同時に選択することができる。

3.7 ノード選択と再マップ表示

自己組織化マップパネルを用いたクラスタリングでは，10×10 の各ノードに必要・普通・不要のいずれかの属性を割り当てて分類する。ノードに含まれる全ての関連キーワードは，ノードと同じ属性が割り当てられる。各ノードを右クリックすると属性を設定するためにポップアップメニューが現れ，メニューの中から属性を選択することでノードに属性を割り当てることができる。ノードに属性を割り当てると，ノードの背景色が必要属性ならば赤色に，不要属性なら青色に，普通属性なら灰色（元の色）に変化する。各ノードに属性を割り当てた時の例を図４に示す。

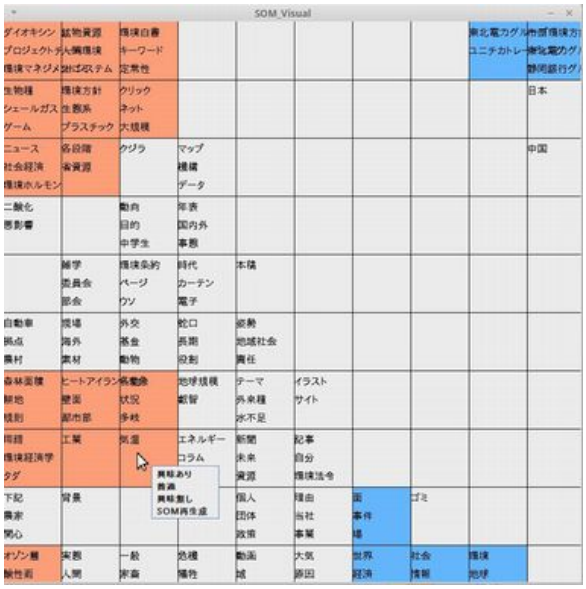

図４　ノードへの属性設定

関連キーワードの分類が完了したら最後に自己組織化マップの再作成を行う。再作成は右クリックで現れるポップアップメニューの最下部にある「再作成」を選択することで実行される。再作成が実行されると不要属性が割り当てられた関連キーワードが削除され，必要・普通属性の関連キーワードだけで再度特徴ベクトルが作られる。このとき特徴ベクトルに新たな属性が 1 つ追加される。追加された属性の値は，必要属性であれば 300，普通属性であれば 0 となる。新たな特徴ベクトルを基に再び自己組織化マップを作成して表示する。このように自己組織化マップの作成とユーザーによる関連キーワードの選別を繰り返し行うことで，ユーザーが興味を持つ関連キーワードや Web ページを絞り込んでいくことができる。

４．実験と考察

4.1 実験方法

本システムにおけるユーザーの関連キーワードの選別から自己組織化マップの再作成・表示までを 1 サイクルと考えたとき，1 サイクルで検索結果から抽出した関連キーワードと，それらを含む Web ページの数をどの程度絞り込んでいくことができるかを調べる。これによりシステムの ExploratorySearch 支援の性能を評価する。実験の条件として，検索数を 100 ページ，使用する Web ページの情報をスニペットとする。以下の表 1 に実験時の検索キーワードと，検索結果の中で必要とした話題の基準，関連キーワードの例を示す。

表 1 実験の ExploratorySearch

検索 キーワード	必要な 話題	必要な キーワード例
環境問題	環境問題の種類	地球温暖化
宇宙	航空技術	ロケット
プログラミング	言語の種類	java
国内旅行	観光地	北海道
海外旅行	観光地	台湾

4.2 実験結果

表 1 に示す検索キーワードで実験を行ったときの，初期の検索結果のキーワード数と Web ページ数を表 2 に示す。ここで，Web ページの数が 100 以下のものがあるが，これは抽出するべき関連キーワードをスニペット中に含まない Web ページが存在したためである。

表 2　初期の数

検索キーワード	キーワード数	Web ページ数
環境問題	482	100
宇宙	585	99
プログラミング	410	100
国内旅行	443	100
海外旅行	436	99
合計	2356	498

1 サイクル，2 サイクル後のキーワード減少率とページ減少率を表 3，表 4 に示す。 キーワード減少率とページ減少率は，以下の式で定義した。

$$\text{キーワード減少率} = \left(1 - \frac{\textit{現在キーワード数}}{\textit{初期キーワード数}}\right) \times 100$$

$$\text{ページ減少率} = \left(1 - \frac{\textit{現在ページ数}}{\textit{初期ページ数}}\right) \times 100$$

表 3　1 サイクル後の減少率

検索 キーワード	キーワード 減少率	ページ 減少率
環境問題	21	0
宇宙	41	2
プログラミング	29	3
国内旅行	87	40
海外旅行	58	7
平均	49	11

表 4　2 サイクル後の減少率

検索 キーワード	キーワード 減少率	ページ 減少率
環境問題	56	11
宇宙	69	13
プログラミング	55	23
国内旅行	90	43
海外旅行	75	43
平均	69	25

4.3 考察

表 3 を見ると 1 サイクル後の自己組織化マップでキーワード減少率の平均は 49%である。1 サイクル目のキーワード選別では必要・普通属性に割り当てられたノード数の合計が 116，不要属性に割り当てられたノード数の合計が 142 とおおよそ同数である。表 4 を見ると，2 サイクル後の自己組織化マップではキーワード減少率の平均が 69%と 1 サイクル後からさらに半減近く減少している。2 サイクル目のキーワード選別では必要・普通属性に割り当てられたノード数の合計が 54，不要属性に割り当てられたノード数の合計が 39 であり，こちらもおおよそ同数である。自己組織化マップの再作成では，不要属性の関連キーワードが削除される。表 3，表 4 の関連キーワードの減少率の平均と，不要属性に割り当てられ削除されたノードの比率がほぼ同じであることから，自己組織化マップの各ノードには関連キーワードがほぼ均等に配置されており，関連キーワードの減少数は不要属性に割り当てるノードの数に比例すると考えられる。したがって，本システムは関連キーワードの絞り込みについては非常に効率よく行うことができると考えられる。

表 3 を見ると 1 サイクル後の Web ページの減少率の平均は 11%であり，キーワードの減少率と比較すると非常に低いことが分かる。また減少率の平均が 11%となったのは検索キーワード「国内旅行」で特

に減少率が多かったためであり，本来の１サイクル後の Web ページ減少率は 11%よりも低い数字だと予想される。表４を見ると２サイクル後でも初期表示からの Web ページ減少率は 25%であり，キーワードの初期表示からの減少率 69%と比較しても低いことが分かる。したがって本システムではキーワード選別によるクラスタリングのサイクルを繰り返しても，キーワードを含む Web ページを絞り込むのは難しいと考えられる。このような結果となった原因は，キーワードに一般的に使用される単語が含まれてしまったことや，必要とした話題の関連キーワードが，検索キーワードを説明する際に一般的に使われるキーワードであったためなどが考えられる。本システムが効率的に Web ページを絞り込めるようにするには，一般的に使用されるキーワードを削除することや，必要とする話題を検索キーワードの中でもマイナーなものにする必要があると考えられる。

５．おわりに

本研究では，Exploratory Search を支援するために検索結果のからの関連キーワード抽出と，ユーザによるキーワード選別およびクラスタリング機能を備えた提案システムの開発，および提案システムの性能を検証するための評価実験を行った。実験の結果から，本システムは検索キーワードに関連した特定の話題のキーワードの絞り込みについては効率よく行うことができるが，それらの話題について詳しく調べるために閲覧する，関連キーワードを含む Web ページの数はクラスタリングを繰り返し行っても大きく減少することはなく，効率よく絞り込むのは難しいことが分かった。Web ページを効率よく絞り込んでいくには，関連キーワードのクラスタリング方法や，必要とする話題の選択を工夫する必要がある。

本研究の今後の課題としては，それぞれ異なるユーザーにとって最適な検索およびクラスタリング結果を得られるようにするために，システムの設定をユーザーが自由に変更できるようにすることが挙げられる。本システムはプログラム側がユーザーに一方的にクラスタリングの結果を提供するのではなく，ユーザー側が主体となって独自の基準でクラスタリングを行うことで Exploratory Search を進めることを目標としている。今回説明したシステムの内容ではユーザーが決定できるのは，検索キーワード，検索ページ数，解析に使用する情報の種類のみであり，それ以外の設定は変更することができない。しかし本来は，特徴ベクトルの属性内容や値，キーワードの重要度の決定方法や計算式の係数の重み，抽出する関連キーワードの品詞の種類，自己組織化マップパネルのノード数など多くの項目をユーザーが自由に変更することで，ユーザー独自のクラスタリングを提供するシステムが理想である。今後はこのような課題を解決するためにシステムの改良を行う必要がある。

参考文献

[１] Jesse Alpert，Nissan Hajaj：We knew web was big…OfficialGoogleBlog-http://www.googleblog.blogspot.jp/2008/07/weknew-web-was-big.html.

[２] RyenW.White,ResaA.Roth:ExploratorySearch:BeyondtheQuery-ResponseParadigm-http://www.morganclaypool.com/doi/abs/10.2200/s00174ed1v200901icr003.

[３] JAKOB NIELSEN:How Long Do User Stay Web Pages?NielsenNormanBlog-http://www.nngroup.com/articles/howlong-do-users-stay-on-web-pages/.

[４] 梶並知記，高間康史：ユーザ意図を強調したキーワード配置支援機能を備えたインタラクティブなキーワードマップ，情報処理学会論文誌，Vol.48，No3，pp.1176-1185，2007.

[５] 津高新一郎：自己組織化マップを用いたテキスト自動分類の試み，情報処理学会 第 46 回全国大会講演論文集，pp.187-188，1993.

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第10回)
SIG-AM-10-04

同義語判定問題を用いた語義ベクトルの評価の検討
ーSkip-gram モデルで獲得した語義ベクトルを例としてー

Evaluation of Word Vectors by Synonym Identification
- Skip-gram Word Vectors as an Example -

城光 英彰[1]　松田 源立[1]　山口 和紀[1]

Hideaki Joko[1], Yoshitatsu Matsuda[1], Kazunori Yamaguchi[2]

[1]東京大学総合文化研究科

[1]Graduate School of Arts and Sciences, the University of Tokyo

Abstract: Automatic synonym acquisition is an important problem in the field of document retrieval and data mining using natural language data. In this paper, we conducted two experiments to identify the properties of word vectors acquired by the Skip-gram model, related to the synonym identification. In the first experiment, we confirmed that the cosine similarity of a synonym pair is significantly higher than that of a non-synonym pair. In the second experiment, we show that only a limited number of components of word vectors are needed for discriminating synonym pairs from non-synonym pairs.

1 はじめに

自然言語処理において人間のような意味処理を実現する上で，言い換え表現の獲得は中心的な課題とされている[1]．そのような言い換え表現獲得を含む汎用な意味処理を実現する一つのアプローチとして，意味の基本単位である「単語の意味」について着目することは有用であると考えられる[2]．例えば，文書検索において「東京大学」を検索する際に，「東京大学」だけではなく「東大」や「UT」などを含めた文書も検索対象としたいと考えている場合，「東京大学」の同義語として「東大」や「UT」を獲得しておく必要がある．また，Web 上の文章を用いたデータマイニングなどにおいても，同じ意味を表す単語の違いにより生じるデータスパースネスを解消する上で，同義語の獲得は重要な課題となる．人手により同義語辞書を作成するアプローチも考えられるが，次々と生まれる新語に対応することが困難であることなど問題点が多く，人手による網羅的な同義語辞書の作成は現実的ではない．このような理由から，同義語の自動推定は重要な課題であると考えられる．

同義語の自動推定には様々な手法が存在する．笠原ら[2]は，国語辞典を用いて，

1. 見出し語に対して語義文より特徴行列を作成する．
2. 特徴行列から大規模なシソーラスを用いて属性行列を作成する．

という処理で属性行列を生成しておき，個々の刺激語が与えられたら，

1. 属性行列の語彙に対して，刺激語と単語親密度が高い語のみを検索対象として絞り込む．
2. 1.で検索対象となった語と属性行列を用いて求めた類似度の高い語を結果とする．

という手法により，刺激語の同義語を推定した．吉田ら[3]は同義語の抽出手法として，

1. コーパスにおいて，検索文字列に隣接する文字列を検索する．
2. 得られた文字列から適切な文脈(文字列)を選択する．
3. 文脈に隣接する文字列を検索する．

という処理により，検索時に実用に耐えうる速度で実行できる同義語の抽出手法を提案している．

これらの手法では，同義語の推定に，何らかの単語の素性を使用している．例えば，[2]では，VSM(ベクトル空間モデル)を使用しており，[3]では隣接文字列を使用している．

「同じ文脈に現れる単語は類似した意味を持つ」という分布仮説(distributional hypothesis)[4]や，実際に文脈情報が同義語判定に有用であるとの報告 [5]から，同義語判定においては文脈情報を活用するこ

とは重要であると考えられる．一方で，吉田ら[3]のように単語の出現頻度のみを用いた VSM は，語順を無視し(必然的に，周辺文脈を無視し)ているが，特異値分解などの手法が適用可能であり，スパースネスなどの問題を緩和できるという利点を持つ．

近年，分布仮説に基づきニューラルネットワーク的な手法を用いて単語の”意味”を表すベクトル(語義ベクトル)を求める Skip-gram モデルが提案された[6]．Skip-gram モデルで得られた語義ベクトルは，加法構成性(後述)を持つことやコサイン類似度により単語の意味の類似度が計算できることが報告されている．Skip-gram モデルで求めた語義ベクトルは，VSM と異なり周辺文脈を考慮に入れており，その語義ベクトルを用いれば，同義語を，従来手法より高精度に判定できる可能性がある．しかし，Skip-gram モデルで求めた語義ベクトルの性質については定量的な分析も少なく，どのように利用すれば同義語の判定に効果的に利用できるかは明らかになっていない．そこで，我々は Skip-gram モデルで求めた語義ベクトルの性質を明らかにするためにいくつか実験を行った．本論文では，その実験の結果を報告する．

本論文は以下のような構成となっている．2 節では Skip-gram モデルについて説明する． 3 節では今回行った実験の内容とその結果を述べる．最後に 4 節でまとめと今後の課題を示す．

2 Skip-gram モデル

ここでは Skip-gram モデル[7]について概説する．まず基本的なモデルについて述べ，次にその近似である階層的 softmax モデルについて述べる．最後に，現在までに報告されている特徴を述べる．

Skip-gram モデルは，ニューラルネットワーク的な手法を用いて，コーパスの文脈情報から，各単語の語義ベクトルを学習する手法の一種である．Skip-gram モデルでは，ある単語w_tが文章内の位置 t に存在した場合，その周囲の単語 w_{t+j} $(j \neq 0)$の発生確率$\mathrm{p}(w_{t+j}|w_t)$を以下の式で与える．

$$\mathrm{p}(w_{t+j}|w_t) \propto \exp\left({v'_{w_{t+j}}}^{T} v_{w_t}\right)$$

ここで，ニューラルネットワークモデル的に言えば，v_wはある入力単語wに依存した入力用ベクトル，v'_wはある周辺単語wの出力確率を計算するための出力用ベクトルである．出力確率は，入力用ベクトルと出力用ベクトルの内積に依存し，内積が大きい程確率は高くなる．本論文では，わかりやすさのため，v_wを単語の語義ベクトル，v'_wを文脈ベクトルと呼ぶことにする．なお，確率分布は 1 に正規化されるので，語彙に含まれるすべての単語wでの正規化により，$\mathrm{p}(w_{t+j}|w_t)$は以下で与えられる．

$$\mathrm{p}(w_{t+j}|w_t) = \frac{\exp\left({v'_{w_{t+j}}}^{T} v_{w_t}\right)}{\sum_w \exp\left({v'_w}^{T} v_{w_t}\right)}$$

さらに$\mathrm{p}(w_{t+j}|w_t)$から，あるコーパスが与えられたときの尤度関数ℓを以下のように定義する．

$$\ell = \sum_{t=1}^{T} \sum_{-c \le j \le c,\ j \neq 0} \log \mathrm{p}(w_{t+j}|w_t)$$

ここで T はコーパスのサイズ，c は事前に与えられる文脈窓のサイズである．実際のコーパスを利用して，ℓを最大化する語義ベクトルv_wおよび文脈ベクトルv'_wを求めることが，Skip-gram モデルにおける学習である．

本来のモデルは上記の通りであるが，尤度関数ℓをこのままの形で最大化することは，計算量等の問題で困難であるため，実際にはいくつかの近似が用いられる．ここでは[7]で採用されている近似である階層的 softmax モデルについて述べる．階層的 softmax モデルでは，文脈の計算において，まず単語群を事前に二分木構造に整理しておく．二分木構造としては様々な候補がありうるが，実験的には，頻度に基づく手法である Huffman 木が有効であることが知られており，[7]でも Huffman 木が用いられている．二分木完成後，文脈ベクトルを，各分岐ノードのみに割り当てる．木構造の葉に相当する実際の周辺単語には文脈ベクトルは割り当てられず，これにより推定すべきパラメータ数は大幅に減少する．あるノード k とある単語 w が与えられた場合，そのノードで分岐の右(right)と左(left)を辿る確率を以下で定義する．

$$\mathrm{p}(\mathrm{right}|\mathrm{k},\mathrm{w}) = \frac{1}{1+\exp(-{v'_k}^{T} v_w)}$$

$$\mathrm{p}(\mathrm{left}|\mathrm{k},\mathrm{w}) = \frac{1}{1+\exp({v'_k}^{T} v_w)}$$

ここで，ある周辺単語が与えられたとすると，二分木内にはその単語に辿りつく唯一のパスが存在する．そして，そのパスは，根ノードから，葉に辿りつくまでに，順番に左右どちらを選ぶかで表現される．従って，そのパスを辿る確率は，$\mathrm{p}(\mathrm{right}|\mathrm{k},\mathrm{w})$もしくは$\mathrm{p}(\mathrm{right}|\mathrm{k},\mathrm{w})$をパスに沿って積算することによって与えられる．以上が階層的 softmax モデルによる近似の原理である．実際の学習においては，確率的勾配法に基づくバックプロパゲーションアルゴリズムが利用される．最大化を効率的に行うため，それ以外にもサンプリング等でいくつかの技法が用いられている．学習アルゴリズムの詳細は[7]を参考にさ

れたい．

最後に，Skip-gram モデルの特徴について述べる．Skip-gram モデルにより学習される単語の語義ベクトルは，単語間の何らかの関係性を学習したものであることが報告されている[7]．具体的には，単語”Berlin”の語義ベクトルから，”Germany”の語義ベクトルを減じ，”France”の語義ベクトルを加算すると，”Paris”の語義ベクトルと近いベクトルとなることが知られている．これは，「国家」の「首都」という単語間の関係が，語義ベクトル空間の中で，加法的な関係として抽出されていることを示しており，加法構成性と呼ばれている．このような関係性が語義ベクトルに埋め込まれていることから，単語の関係の主要なものの一つである「同義関係」も語義ベクトルに埋め込まれていることが期待される．しかし，Skip-gram モデルの実験的/理論的な性質には未解明の部分が多いため，本論文では「同義関係」に絞って挙動の詳細な分析を行う．

3 実験

3.1 実験目的

本実験の目的は同義語推定に関係する，Skip-gram モデルの性質を明らかにすることである．そのために，ここでは 2 種類の実験を行った．

実験 1

Skip-gram モデルに関する研究では，意味の類似度を語義ベクトルのコサイン類似度で測るのが一般的である([6], [8]). 同義語は意味の”距離”が近い単語であることから，同義対のコサイン類似度は非同義語対と比較して高い値をとるはずである．そこで，実験 1 では，同義対のコサイン類似度が非同義語対のものよりも高いことを確認する．同義対のコサイン類似度の値には幅があることが予想されるため，特定の閾値を設けて同義語か非同義語かを判定してその精度を見るのではなく，類似度の分布を用いて比較を行った．

実験 2

上で述べたように，語義ベクトルのコサイン類似度により意味の類似度を測る事が一般的に行われているが，語義ベクトルの全ての成分が意味の距離に等しく影響を与えているとは考えにくい．そこで実験 2 では，同義語推定においてベクトルの各成分が与える影響の違いを調査した．なお，実験には線形 SVM を使用したが，これは各成分が与える影響の解釈しやすさを考慮してのことである．

3.2 データ

語義ベクトル作成において用いたコーパスとしては，日本語 Wikipedia データ[1](2Gbytes)を MeCab[2]を用いて基本形出力でわかち書き行った後に，出現回数が 30 万回以上の高頻度語と 100 回未満の低頻度語を除いた 78274 語を使用した．Skip-gram[3]モデルでは，語義ベクトル(および文脈ベクトル)の次元は 200，文脈の広さ c は 5 として，階層的 softmax モデルを用いて学習を行った．求められた語義ベクトルには L2 ノルムが 1 となるよう正規化処理を施してから実験 1 と実験 2 で使用した．

実験 1 及び実験 2 において，同義語の教師例としては Wordnet 同義対データベース[4]を使用した．同義語データベースの中で語義ベクトルが獲得されている 8373 対を正例とし，語義ベクトルが獲得されている 78274 単語の中からランダムに選んだ 8373 対を負例とした．実験 2 においては，これらの単語対に対応する 200 次元の語義ベクトル 2 つを，論文[9]を参考にして，図 1 のように結合し 400 次元のベクトルとして線形 SVM を適用した．論文[9]では，ベクトルの成分毎の和や差を用いた分析も行っているが，結合したベクトルは，和や差のベクトルよりも情報が多く，元のベクトルとの要素の対応も単純で解釈しやすいことから，本実験では結合したベクトルを用いた実験のみを行った．

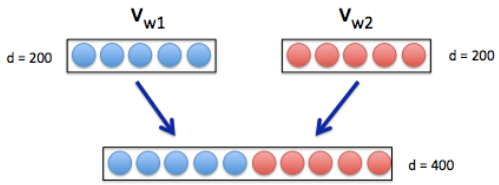

図 1: 実験 2 で用いたベクトルの構成

[1] http://dumps.wikimedia.org/jawiki/ (accessed 2015-5-12).
[2] http://taku910.github.io/mecab/ (accessed 2015-5-29).
[3] https://code.google.com/p/word2vec/ にて Google が公開しているものを使用した．(accessed 2015-2-15).
[4] http://nlpwww.nict.go.jp/wn-ja/jpn/downloads.html にて NICT が提供する，Wordnet[10]を元に作成された同義対データベースである．(accessed 2015-6-6).

3.3 実験 1:コサイン類似度の分布

実験 1 では正例と負例のコサイン類似度の分布の比較を行った．結果を図 2 に示す．平均値については，正例では 0.258，負例では 0.075 となった．正例と負例の分布について，等分散の仮定の下で，右片側 2 標本 t 検定を適用したところ，t 統計量は 71.7 となり，p 値はほぼ 0 であった．これにより，正例と負例でコサイン類似度が有意に異なることが確認できた．

表 1 にコサイン類似度が上位 10 位の同義対, 及び下位 10 位の同義対を示す. コサイン類似度が上位である同義対と比較すると，下位である同義対には一方の単語が複数語義を持つもの(例えば，サークル，ポイント)が多く見受けられる．また，下位 10 例の同義対においては, すべて, 一方が外来語(カタカナ)であり，もう一方が和語や漢語(平仮名や漢字)である．逆に，類似度の上位 10 例は，外来語同士か和語/漢語同士が対になっていることから，外来語と和語/漢語は同義であっても周囲に異なった単語分布を生む可能性が高いと考えられる．

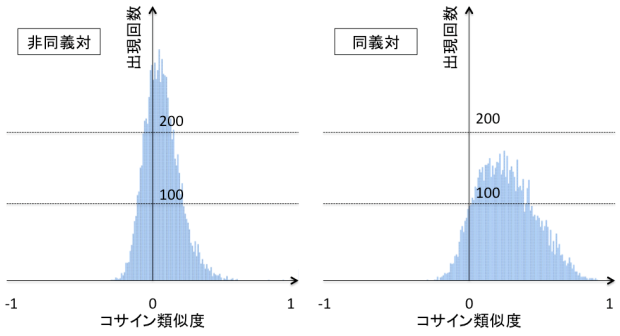

図 2: 同義対(右)と非同義語対(左)の
コサイン類似度の分布

3.4 実験 2:SVM による分類

実験 2 として線形 SVM を用いて同義語の判定を行った．ここでは，分類の性能を上げることが目的ではなく，語義ベクトルの各成分の影響を明確にすることが目的であるため，線形カーネルを用いた．10 分割交差検定により求めた分類の結果は，正答率が 92.59%，精度と再現率は 0.913 と 0.942，F 値は 0.927 となった．また，Confusion matrix を表 2 に示す．これらの結果より，一定程度の分類性能を持っていることが分かることから，意味のある超平面が構成されていると考えられる．

表 1: 同義対とそのコサイン類似度
(上位 10 例と下位 10 例)

同義対	コサイン類似度
ウェブブラウザ, ブラウザ	*0.88946885*
相打ち, 相討ち	*0.87708294*
サイト, ホームページ	*0.8724321*
ウェブサイト, サイト	*0.87111324*
反乱, 叛乱	*0.86343205*
敵意, 敵愾心	*0.85473984*
ウェブサイト, ホームページ	*0.8524012*
吃水, 喫水	*0.8515739*
憤慨, 激怒	*0.8491848*
考え, 考え方	*0.8447531*
・	・
・	・
・	・
キャリヤ, 経歴	*-0.19684665*
サイン, 兆	*-0.1994007*
キー, 緒	*-0.21134022*
サム, 和	*-0.21272291*
サークル, 丸	*-0.21344633*
ルール, 定則	*-0.21581507*
ノース, 子	*-0.21692836*
ポイント, 地	*-0.22948295*
ハイム, 家作	*-0.2397215*
ラック, 幸	*-0.2841633*

次に線形 SVM により算出された 400 次元の各成分の重みの度数分布を図 3 に示す．図 3 から，単語ベクトルの全ての成分が同義語決定に等しく影響を与えているのではなく，少数の成分が大きな影響を与えていることが分かる．

表 2: Confusion matrix

		予測されたクラス	
		正例	負例
実際のクラス	正例	7888	485
	負例	756	7617

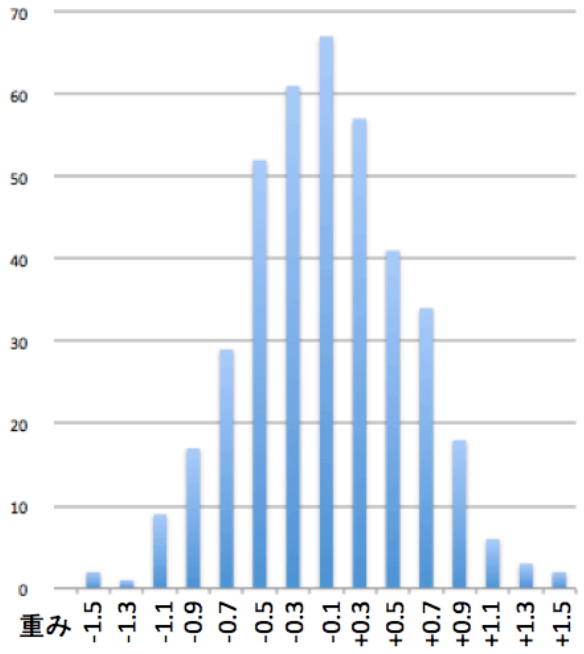

図 3: 線形 SVM の重みの度数分布

4 まとめと今後の課題

本論文では，同義語推定への応用を視野に入れ，Skip-gram モデルを用いて得られた語義ベクトルの性質を明らかにするための 2 つの実験を行った．実験 1 では，同義語のコサイン類似度が非同義語のコサイン類似度より有意に高いことが確認された．同義語であるにもかかわらずコサイン類似度が低いものには，どちらか一方の単語が複数語義を持つものや，一方が外来語で他方が和語/漢語であるものが見られた．単語が複数語義を持つ問題は，単語のもつ語義ごとに語義ベクトルを生成する Skip-gram モデルの拡張(例えば，[11])などを用いることで解決できる可能性がある．また，外来語と和語/漢語の問題は，外来語自体はどの言語にもあるが，今回用いた Wordnet の同義語対データベースが英語の同義語対の翻訳であることに起因する可能性がある．この点については英語の Wordnet で実験することで日本語 Wordnet や翻訳による影響がどの程度あるかを検討することができるであろう．実験 2 の結果からは，語義ベクトルには同義語判定において重要な役割を果たす成分と，それほど影響を与えない成分が存在することが明らかとなった．今後は，同義語の語義ベクトルと線形 SVM の重みの対応を検討することで，各成分の担う意味を明らかにし，これらを進めることで，同義語の推定手法を構築していきたい．

参考文献

[1] 乾健太郎: 自然言語処理と言い換え，日本語学，Vol. 26, No. 11, pp. 50–19, (2007).

[2] 笠原要，稲子希望，加藤恒昭: テキストデータを用いた類義語の自動作成，人工知能学会論文誌，Vol. 18, No. 4, pp. 221-232, (2003).

[3] 吉田稔，中川裕志，寺田昭: コーパス検索支援のための動的同義語候補抽出，人工知能学会論文誌(Web), Vol. 25, No. 1, pp. 122-132, (2010).

[4] Harris Zellig: Distributional structure, Word, Vol. 10, No. 23, pp. 146-162, (1954).

[5] Hagiwara, Masato, Yasuhiro Ogawa, Katsuhiko Toyama: Selection of Effective Contextual Information for Automatic Synonym Acquisition, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp. 353–360, (2006).

[6] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space, Workshop at International Conference on Learning Representations (ICLR), (2013)

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean: Distributed Representations of Words and Phrases and their Compositionality, Advances in Neural Information Processing Systems (NIPS), (2013).

[8] 椿真史, Kevin Duh, 新保仁，松本裕治: 文の意味構成に伴う高次元空間の最適化と単語表現学習，言語処理学会第 20 回年次大会発表論文集，pp. 1015-1018, (2014)

[9] Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, Bill Keller: Learning to Distinguish Hypernyms and Co-Hyponyms, Proceedings of the 25th International Conference on Computational Linguistics, (2014).

[10] Francis Bond, Timothy Baldwin, Richard Fothergill, Kiyotaka Uchimoto Japanese SemCor: A Sense-tagged Corpus of Japanese, The 6th International Conference of the Global WordNet Association (GWC-2012), Matsue, (2012).

[11] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, Andrew McCallum: Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space, EMNLP, (2014).