Vol.2017-NL-231 No.26
Vol.2017-SLP-116 No.26
2017/5/16

文書構造を考慮した部分文書検索手法の提案

本間 幸徳[1]　貞光 九月[1]　西田 京介[1]　浅野 久子[1]　松尾 義博[1]

概要：本稿では，ある文書におけるユーザの検索要求に対し，一つ以上の文を回答として提示する部分文書検索手法を提案する．検索要求によっては提示すべき文が文書中に散在する場合があるため，提案手法では，文間の関係性に基づいて推定した文書構造を用いることで，文書に散在する文の集合を部分文書として抽出する．また抽出された部分文書について，分散表現を利用した意味ベクトルを作成し，検索スコアの算出に用いることで検索精度の向上を図る．評価実験により，文書構造に基づいて部分文書を抽出し，対応する意味ベクトルを検索に用いることで，ユーザの検索要求に適した検索結果が得られることを示す．

Partial Document Retrieval based on Document Structure

Yukinori Homma[1]　Kugatsu Sadamitsu[1]　Kyosuke Nishida[1]　Hisako Asano[1]
Yoshihiro Matsuo[1]

1. はじめに

ウェブ上の文書の増加に伴い，文書の中からユーザの求める情報を探し出す検索技術の重要性が高まっている．多くの Web 検索システムでは，ユーザの検索要求に対して文書全体を回答として提示するが，例えば商品の説明書や約款など，比較的長い文書においては，文書中の各部分によって記述される情報が異なる場合がある．これらの比較的長い文書では，ユーザは自身が求める情報を文書全体から探し出す必要があり，検索時の負担となっている．

本稿で取り組む部分文書検索技術は，ある HTML 文書におけるユーザの検索要求や質問に対して，最も良く適合する文集合 (部分文書) を文書の中から抽出し，回答として提示する検索技術である．部分文書検索を利用することで，文書中から回答の記述部を探すための時間的なコストが小さくなるため，検索時におけるユーザの負担軽減に効果があると考えられる．

部分文書検索では，どのように部分文書を抽出するかが一つの課題となる．回答として提示する部分文書は，ユーザの検索要求に対して必要な文を含み，無関係な文を含まない文集合から構成されることが望ましい．文書からユーザの検索要求に適合する部分文書を抽出する手法として，例えば，文書中の段落や XML 文書におけるタグ情報に基づいて部分文書を抽出する方法 [1], [10] や，検索キーワードに関する単語・フレーズを含む連続する文集合を部分文書として抽出する手法 [3] がある．

一方で本稿で検索対象として扱う HTML 文書において，前述した手法では検索要求に適合する部分文書が抽出できない場合がある．HTML 文書は閲覧者に対する可読性を高めるために，見出し構造やリスト構造・テキストのフォントサイズや色の変更等を示す HTML タグが数多く含んでいる．そのため文書の物理的な構造と意味的な構造の間に整合性が取れていない場合が多く，XML 文書に対する検索技術をそのまま適用しても高い精度を得られないと指摘されている [2]．

また文書によっては，補足的な情報を示す文が文書中に散在している場合やリスト構造において検索要求に対して無関係な文を含む場合など，提示すべき文集合が連続していない場合がある．このような文書から検索要求に対して適合する部分文書を抽出するためには，HTML タグによる物理的な文書構造だけでなく，文書の意味的な構造を考慮する必要がある．

上記課題を踏まえ，本稿では HTML 文書に対する新たな部分文書検索技術を提案する．提案する手法では，まず

[1] NTT メディアインテリジェンス研究所
1-1, Hikarinooka Yokosuka-Shi, Kanagawa 239–0847, Japan

文書中の文間における意味的な階層関係に基づいて HTML 文書の意味的な構造を推定する．推定した文書構造に基づいて，検索要求に適合する部分文書を抽出・検索する．

また，近年では単語の意味情報として，大規模なコーパスから学習した単語の分散表現を用いる手法 [4] が提案されている．単語の分散表現を用いることで，表層のキーワードだけでなく意味的な類似度を扱うことができるため，文の類似度計算 [5] や情報抽出 [6] 等のタスクにおいて利用され，精度の向上が報告されている．

本手法においても，抽出した部分文書に対して分散表現を用いた意味表現ベクトルを算出し，検索文から生成した意味表現ベクトルとの類似度を検索スコアとして用いることで検索精度の向上を図る．

本稿の貢献は以下の通りである．

- HTML 文書に対するユーザの検索要求に適した部分文書を抽出するために，文間の意味的な階層関係に基づく文書構造の推定手法を新たに提案する．
- 部分文書検索のために，分散表現を利用した質問文・回答テキスト間の類似度比較手法を適用する．評価実験により検索精度が向上することを示す．

2. 関連研究

本節では，ユーザの検索要求に対する部分文書の検索に関連する従来研究について述べ，本研究の位置づけを説明する．部分文書検索に関連する技術として，スニペット [7] やパッセージ検索技術 [1], [3], [8], [9]，XML 部分文書検索技術 [10], [11] などがある．

スニペットは，多くの Web 検索システムにおいて提示される，クエリキーワードとその周辺のテキストからなる要約文である．システム利用者が，検索結果中からいずれの文書が閲覧するのに適切であるかを判断するために用いられ，主にクエリキーワードに基づいて，一つ以上の文やフレーズを一つのテキストとして提示する．スニペットで抽出されるテキストはクエリキーワードを含むテキスト集合であり，文書中の物理的な構造や意味的な構造等に基づく意味的な関連については考慮していない．そのため，検索要求中の語を含まないテキストは提示することができず，また検索要求には直接関連しないテキストを提示する場合も存在する．一方で提案手法は分散表現を考慮することで，検索クエリのキーワードだけでなく類似する語を含む部分文書を抽出することができる．また，文書の意味的な構造を推定しているため，検索要求に対して無関係な文を含まない部分文書を抽出できる可能性がある．

パッセージ検索は，検索クエリに基づいて文書中の連続している文集合を抽出し回答として提示する検索技術である．抽出手法として，文書の形式段落を利用する手法 [1] や，固定のウインドウ幅を設定する手法 [8]，同一のトピックで記述された範囲を特定する手法 [3], [9] などが提案されている．パッセージ検索では，文書中の連続している文集合を検索の対象としているが，提案手法では，文書の意味的な構造に基づいて部分文書を抽出しており，文書中に散在する文集合を回答として提示できるため，より多様な部分文書を検索対象としている．

XML 検索は，マークアップ文書の一つである XML 文書に対する検索技術である．Wikipedia 記事や IEEE 論文に関する XML 文書データに対して，XML タグに基づく文書の木構造における部分木を部分文書として抽出する手法が提案されている [10], [11]．XML 構造は文書の意味的な構造と物理的な構造が一致しているため，ユーザの求める検索要求に対して意味的に関連のある回答を提示することができる．一方で，HTML 文書は文書の物理的な構造と意味的な構造が一致していない場合が多いため，文書の物理的な構造が意味的な構造と一致していることを前提とする，XML 部分文書検索技術を HTML 文書に適切に適用するのは難しい．

HTML 文書に対する XML 部分文書検索技術を適用するための手法として，欅ら [2] は HTML の階層情報を示す特定のタグに基づいて，文書の意味的な階層構造を推定する手法を提案している．ユーザの検索要求に対して，文書の意味的な構造を考慮しない場合に比べ，適切な範囲の部分文書を抽出できることを報告しているが，特定のタグに関連しない文に対して文書構造を推定できない点に工夫の余地がある．一方で，提案手法は上記手法で用いられていないタグ情報や文中の意味情報を考慮して文書構造を推定することで，上記手法では獲得できなかった部分文書を獲得できる可能性がある．

3. 提案手法

提案手法では，文書中の二文間における意味的な階層関係を考慮することで文書構造の推定を行う．推定した文書構造に基づいて，意味的に関連のある部分文書を抽出する．抽出された部分文書において，ユーザの検索要求に対する検索スコアを計算し，回答として適切な部分文書を提示する．

3.1 文書構造の推定

本節では，HTML 文書中の二文間における意味的な階層関係の有無に基づいて，文書構造を推定する方法について説明する．

一般に，文章の論理的な構造は，章，節，項もしくは段落など階層的に分割された部分から構成される [12]．HTML 文書においては，構造を表すタグである H タグや BODY タグ，P タグなどの HTML タグを用いて文書構造が表現されており，特に，H1-H6 の Heading タグは，内容の重要度に合わせて Heading のレベル (タグ中の数値) が設定されるために，階層的な文書構造と見做すことができる．

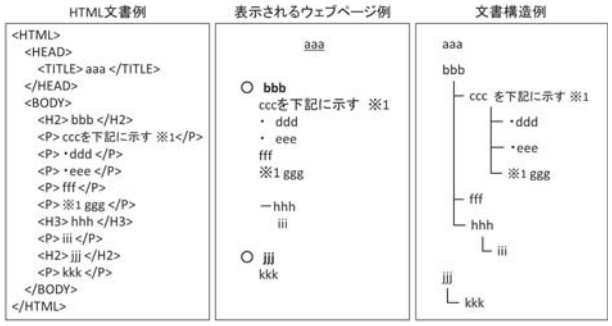

図 1 HTML 文書と文書構造の例
Fig. 1 An example of documents and document structures.

欅ら [2] は Heading タグの情報を利用し，Heading タグの重要度に基づいて文書のの階層的な構造の一部を推定している．

一方で HTML 文書中には，Heading タグは用いずに意味的な階層関係を持つテキストのペアが存在する．図 1 の左に HTML 文書の一例を，中央にレンダリングされたウェブページの一例を示す．例えば，文「ccc を下記に示す ※ 1」と文「・ddd」のペアのように HTML 文書上で用いられている HTML タグは同一である一方で，リスト構造に関する見出しと要素といった階層関係を持つペアや，文「ccc を下記に示す ※ 1」と文「※ 1ggg」のように，記号「※ 1」を用いた注釈に関する修飾・被修飾といった階層関係を持つ文のペア等がある．このような二文間においては，HTML タグ等の情報では関係性の有無が判断できない場合があり，特定の単語・フレーズや記号など，文中の意味情報を考慮する必要がある．

本稿では，文書中の HTML タグと，文中の単語情報を利用した下記のルールを定義し，ルールを全て満たす二文間に意味的な階層関係が有ると推定する．

- ルール 1
  対象とする二文は下記の条件を全て満たす
  【1-1】対象とする二文は特定の同一タグ（BODY, SECTION, DIV, TR）内に存在する
  【1-2】対象とする二文の文書中の出現順と，階層関係の順が一致する
- ルール 2
  対象とする二文は下記の条件のいずれかを満たす
  【2-1】対象とする二文が，同一の注釈を示すフレーズ (”※ 1”など) を含む
  【2-2】対象となる二文において，階層が上位の文中に”下記”，”以下の”，”次の”等の単語を含み，かつ階層が下位の文の開始文字が記号である
  【2-3】対象となる二文において，階層が上位の文が Heading タグ (H ”x”) に囲まれており，かつ階層が下位の文が Heading タグ (H ”x+1”) に囲まれている．
  【2-4】対象となる文において，階層が上位の文が特定の HTML タグ (H”x”, LI, TD) に囲まれており，かつ階層が下位の文が特定の HTML タグ (LI, TD, P) に囲まれている．
- ルール 3
  ある文に対して抽出ルール 1，2 を満たす階層が上位の文が複数ある場合，文書中の出現位置が最も近い文との間に階層関係が有ると推定する

具体例を用いて説明する．図 1 の左に示す元となる HTML 文書に対して，ルール 1 の適用により，文「bbb」と文「」ルール【2-4】を適用することにより文「bbb」と文「ccc を下記に示す ※ 1」や文「jjj」と文「kkk」など，HTML タグに基づいて階層関係が有ると推定できる．また，ルール【2-1】やルール【2-2】の適用により文「ccc を下記に示す ※ 1」と文「・ddd」や文「ccc を下記に示す ※ 1」と文「※ 1ggg」など，文中に意味情報を考慮して階層関係の有無を推定することができる．

推定された二文間の階層関係の有無に基づいて，構築される木構造を文書構造として見なす．図 1 の右に推定される文書構造の例を示す．

3.2 部分文書の抽出

本節では，前節で推定した文書構造木に基づいて，HTML 文書から部分文書を抽出する方法について説明する．推定した文書構造について，すべての部分木を部分文書として検索対象とした場合，検索コストが過大となる恐れがあるため、XML 文書構造木からの部分文書抽出手法 [2] を修正することで，HTML 文書への適用を行う．

- ルール 1
  葉ノードに対応する文について，根ノードまでの先祖ノードを順に追加した文集合を一つの部分文書として抽出する
- ルール 2
  子を持つノードに対応する文について，子孫ノード全ての文を含む文集合を一つの部分文書として抽出する

さらに，上記ルールと 3.4 節に後述する部分文書同士の結合を行うことで，ユーザの検索要求に対して必要十分な部分文書を抽出することが可能となる．

図 2 に，図 1 の HTML 文書から抽出される部分文書の例を示す．上述したルール 1 より「部分文書 A」に示す，文「ccc を下記に示す ※ 1」を親として子孫ノードを全て含む部分木に対応する部分文書が抽出される．同様にルール 2 より「部分文書 B」や「部分文書 C」に示す，先祖ノードを含む部分木に対応する部分文書が抽出される．また，3.4 節で述べる部分木の結合処理によって，「部分木 B+C」のように検索要求に適合する部分木を部分文書として抽出することができる．

ここで推定された文書構造によっては，いくつかの部分文書を内包する，文字数が大きい部分文書が抽出されうる

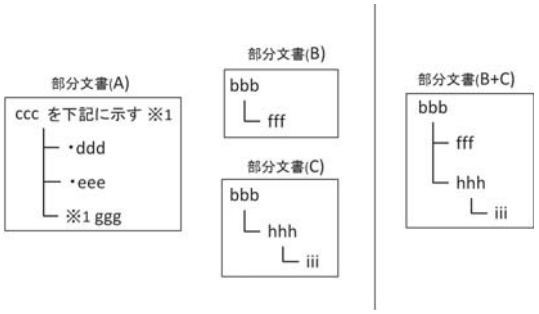

図 2　部分文書例
Fig. 2　An example of partial documents.

が，そのような文書構造はユーザの検索要求に対して非適合箇所を含む可能性が大きいと考え，一定以上の文数を持つ部分文書は抽出の対象外とする．実験では 8 文以上の文数を持つ部分文書を抽出の対象外とした．

3.3　検索スコアリング

本節では，抽出された部分文書に対して，検索要求となる質問文が入力された際の検索スコアリング手法について説明する．

文書検索や XML 検索におけるスコアリングスコアリング手法として，TF-IDF や BM25 などの手法があるが，BM25 を利用した手法が高精度な検索精度を示すことを報告されており [11]，本稿では BM25 をスコアリング手法として利用する．

n 個の単語 q を含む検索クエリ $Q(=\{q_1, q_2, \cdots, q_n\})$ が与えられた時の各文書 D に対する BM25 のスコア S_{BM25} を次に示す．

$$S_{BM25} = \sum_{i=1}^{n} IDF(q_i) \frac{TF(q_i)(k_1+1)}{TF(q_i) + k_1(1 - b + b\frac{|D|}{avgdl})} \quad (1)$$

$$IDF(q) = \log \frac{N - d(q) + 0.5}{d(q) + 0.5} \quad (2)$$

ここで，$TF(q)$ は文書 D 中の単語 q の出現数を，$d(q)$ は単語 q を含む文書数を示す．また，N は部分文書の総数，$avgdl$ は部分文書の平均文字数を示す．また，k_1, b はハイパーパラメタであり，$k_1 = 2.0$，$b = 0.75$ に設定した．

BM25 は表層的な単語情報を用いる手法であり，対象とする文書が検索要求のキーワードを含まない場合適切に検索できない場合がある．本稿では，意味的に近い類似キーワードから検索可能とするために，単語の分散表現を利用するスコアリング手法を利用する．具体的には，単語の分散表現に基づいて質問文と部分文書を示す文ベクトルをそれぞれ算出し，双方のベクトルのコサイン類似度を分散ベクトルによるスコア S_{cos} として用いる．

近年分散表現を利用した文のベクトル算出手法について，単語の分散表現の重み付き平均による手法が単純かつ，効果的であることが報告されており [5]，本稿では，検索対象の文書全体における単語の出現確率 $p(w)$ による重み付き平均ベクトルを文ベクトルとして扱う．

$$s_{vec} = \frac{1}{L} \sum_{j=1}^{L} \frac{\alpha}{\alpha + p(w_j)} v(w_j) \quad (3)$$

ここで L は対象とするテキストに含まれる，名詞・動詞・形容詞などの内容語の総数を示す．また，w_i は部分文書における i 番目の内容語を，$v(w_i)$ は語に対応する単語ベクトルを示す．ここで単語ベクトルとして wikipedia をコーパスとして学習した 200 次元の word2vec[4] を用いた．α はハイパーパラメタであり，本稿では $\alpha = 10^{-3}$ に設定した．

上記の手法に基づいて，質問文と部分文書を文ベクトルに変換後，双方のベクトルのコサイン類似度を分散ベクトルによるスコア S_{cos} として算出する．

BM25 のスコアと文ベクトルのコサイン類似度の線形和を検索に用いるスコアとして算出する．

$$S = S_{BM25} + \beta S_{cos} \quad (4)$$

β はスコアリングにおける係数であり，本稿では $\beta = 4.0$ に設定した．

3.4　部分文書の結合

本手法では，ユーザの検索要求入力時における部分文書の結合手法について説明する．検索要求に対して部分文書の結合を考慮することで，事前にすべての部分木を部分文書として抽出することなく，検索要求に対して必要な文を含む部分木を抽出することができる．

具体的には，検索スコアリングによって算出される検索スコアに基づく部分文書の順序付きリストを考え，最も高い検索スコアを持つ部分文書と次に高い検索スコアを持つ部分文書に対して，以下のルールを全て満たす時に，二つの部分文書の結合を行う．ルールを満たさない場合，最も高い検索スコアを持つ部分文書を除いた順序付きリストに対して，同様の処理を行う．上記処理を繰り返し適用することで，全ての部分文書に対して結合を考慮することができる．

- ルール 1
  二つの部分文書間において一つ以上の文が重複している
- ルール 2
  部分文書の結合により，検索クエリの単語をより多く含む部分文書が抽出される
- ルール 3
  結合により抽出される部分文書の文数が一定数以下である．

4.　評価実験

4.1　データセット

本節では，実験に用いたデータセットについて説明する．

表 1　データセット
Table 1　Dataset.

	文字数	文数	質問数	回答平均文数	回答平均文字数
文書 1	5,190	176	105	4.22	149.8
文書 2	10,262	359	204	4.14	125.4

表 2　アノテートデータ
Table 2　Annotation data.

		作業者 B	
		関係有り	関係無し
作業者 A	関係有り	251	37
	関係無し	29	63,586

表 3　階層関係の推定結果
Table 3　Results of relational extraction.

	適合率	再現率	F 値
文書 1	0.234	0.779	0.360
文書 2	0.331	0.805	0.468

データセットとして，2 種類の保険商品のパンフレットに関する HTML 文書を用いる．

各文書に対して，質問文の作成と，対応する回答となる文集合のアノテートを 2 名の作業者によって実施した．表 1 に各文書の概要を示す．

また，各文書に対して，文書中の全ての二文間における意味的な階層関係の有無のアノテートを 2 名の作業者によって行った．2 名の作業者による階層関係の有無に関するデータを表 2 に示す．2 人の作業者によるアノテートデータにおいて，κ 係数は 0.882 であった．

4.2　階層関係の推定

提案手法による意味的な階層関係の抽出精度を表 3 に示す．表 3 より，提案した階層関係の抽出精度では，高い再現率を示しており，二文間の階層関係の多くを抽出できていることがわかる．

本手法では，階層関係の誤抽出により誤った部分文書が作成されることよりも，階層関係の未抽出により正解となる部分文書の作成されないことの方が，検索の精度の観点でより大きな問題となると考えられる．部分文書の誤作成は課題に残るものの，部分文書の検索における影響は比較的小さいと考え，検索の評価を行った．

4.3　検索結果・考察

本節では下記に示す比較手法に対して，質問文を入力した際に，どの程度正しく回答である部分文書が検索できるか評価した結果を示す．

- 文検索
  文書中の一文ずつを検索対象とする検索手法
- 段落検索
  特定の HTML タグ (DIV，SECTION) を用いて分割されたテキスト (段落) を検索対象とする手法
- 提案手法
  3 節で説明した，自動推定した階層関係に基づく文書構造木を考慮した部分文書検索手法
- 提案手法 (人手)
  アノテートされた階層関係に基づく文書構造木を考慮した部分文書検索手法

表 4　検索精度
Table 4　Results of partial document retrieval.

	@1			@5		
	適合率	再現率	F 値	適合率	再現率	F 値
文検索	0.514	0.240	0.298	0.799	0.413	0.510
＋ svec	0.523	0.248	0.306	0.803	0.419	0.518
段落検索	0.331	0.639	0.366	0.471	0.845	0.511
＋ svec	0.329	0.633	0.363	0.471	0.850	0.511
提案手法	0.432	0.481	0.423	0.585	0.681	0.584
＋ svec	0.451	0.494	**0.439**	0.584	0.692	**0.588**
提案 (人手)	0.470	0.498	0.452	0.636	0.681	0.628
＋ svec	0.483	0.499	**0.457**	0.646	0.705	**0.643**

各質問文に対して，上位 $N(=1,5)$ 件を提示した際の検索精度を表 4 に示す．評価尺度として，部分文書の文字数に関する適合率，再現率及び F 値を用いた．提示される N 件の部分文書において，最良の F 値を示した結果のマクロ平均を示す．

表 4 では，各手法に対して BM25 のみの検索スコアリング手法を用いた場合と，分散表現を利用した検索スコアを併用した場合の結果を示している．表 4 を見ると，文検索手法は，適合率が高い一方で再現率が低いことがわかる．今回用いたデータセットでは回答として複数の文を求める質問文が多く含まれていたため，一つの文を検索する文検索手法では回答となる部分文書を適切に抽出できなかったことを示している．また段落検索手法では，再現率が高く適合率が低い．HTML タグを利用した段落では，検索要求に関係しない文を多く含み精度を下げる要因になったと考えられる．

今回の実験では提案手法が，文検索手法や段落検索手法に比べ高い F 値を示した．また，人手によるアノテートデータを用いた提案手法が全種訪中で最も高い F 値を示した．提案手法により，ユーザの検索要求に対して必要十分な部分文書を提示できることが確認できた．

また，各手法に対して分散表現を利用したスコアを併用することで，段落検索を除く手法において精度向上が確認された．具体例として，例えば「夫は対象となりますか？」という保険の加入に関する検索要求に対して，分散表現を併用した文検索手法では「…，記名被保険者の配偶者」という検索要求に直接記述されていない単語「配偶者」を持つ文が回答として提示された一方で，BM25 によるキーワードを用いた文検索手法では上記の文を回答として提示できなかった．これは，分散表現の利用により，検索要求に含

まれる「夫」という単語と，提示された文中の「配偶者」が近しい意味を持つと推定されたためと考えられる．比較的短いテキストについては上記のような意味情報を用いた改善が見られたが，比較的長いテキストについては顕著な差は見られなかった．単語数の多いテキストでは検索要求として重要でない単語に影響されることが多く，文ベクトルの算出手法については改善の余地があると考えらえる．

5. おわりに

本稿では，HTML 文書に対するユーザの検索要求に対し，一つ以上の文を回答として提示する部分文書検索手法を提案した．提案手法は，まず文書中の二文間における階層関係を考慮することで文書の意味的な構造を推定する．推定した文書構造に基づいて，意味的に関連のある部分文書を抽出・検索することで，ユーザの検索要求に対して適切な部分文書を提示することができる．また，抽出した部分文書に対して分散表現を用いた検索スコアリング手法を利用することで，表層的なキーワードだけでなく文の類似度を考慮した検索を行った．

評価実験により，文書構造に基づいて部分文書を抽出し，対応する意味ベクトルを検索に用いることで，ユーザの検索要求に適した検索結果が得られることを示した．

本稿では，文書構造の推定手法・部分文書の抽出手法についてルールを用いた手法を採用した．評価実験により文書構造の推定と文書構造に基づく部分文書の抽出・検索手法が効果的であることを示した一方で，ルールによる抽出誤りの課題や，検索システム全体の精度向上に向けた課題が明らかとなった，今後はこれらの課題に対して検討を進めていきたい．

参考文献

[1] Gerard Salton, James Allan, and Chris Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 49–58. ACM, 1993.

[2] 欅惇志, 宮崎純, 波多野賢治, 山本豪志朗, 武富貴史, 加藤博一. Xml 部分文書検索技術の web 文書への適用. *DEIM Forum*, 2014.

[3] 望月源, 岩山真, 奥村学. 語彙的連鎖に基づくパツセージ検索. 自然言語処理, Vol. 6, No. 3, pp. 101–126, 1999.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[5] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations. To Appear*, 2017.

[6] Mahesh Joshi, Ethan Hart, Mirko Vogel, and Jean-David Ruvini. Distributed word representations improve ner for e-commerce. In *Proceedings of NAACL-HLT*, pp. 160–167, 2015.

[7] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, Vol. 1. Cambridge university press Cambridge, 2008.

[8] James P Callan. Passage-level evidence in document retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 302–310. Springer-Verlag New York, Inc., 1994.

[9] Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. Unsupervised text segmentation using semantic relatedness graphs. Association for Computational Linguistics, 2016.

[10] 波多野賢治, 渡邉正裕, 吉川正俊, 植村俊亮ほか. 情報検索技術を用いた xml 部分文書の検索手法. 情報処理学会論文誌データベース (TOD), Vol. 42, No. SIG08 (TOD10), pp. 36–46, 2001.

[11] Patrice Bellot, Toine Bogers, Shlomo Geva, Mark Hall, Hugo Huurdeman, Jaap Kamps, Gabriella Kazai, Marijn Koolen, Véronique Moriceau, Josiane Mothe, et al. Overview of inex 2014. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 212–228. Springer, 2014.

[12] 森岡健二. 文章構成法: 文章の診断と治療. 至文堂, 1963.

[13] Yu Xu and Yannis Papakonstantinou. Efficient keyword search for smallest lcas in xml databases. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 527–538. ACM, 2005.