階層的な見出しブロック構造の分析に基づく Web 検索
Web Search Based on Hierarchical Heading-Block Structure Analysis

真鍋 知博♡

Tomohiro MANABE

本論文では **Web** ページ中の見出し構造の抽出と応用について論じる．ここで見出しとは文書の一部分の非常に簡潔な要約であり，見出しのついた一部分をブロックと呼ぶ．ブロックは他のブロックを包含することがあり，あるページ中の見出しとブロックは階層的な見出し構造を成す．本論文ではまず，**Web** ページ中の見出し構造を自動抽出する手法を提案する．抽出のためには，ページ中に同じ見た目の見出しが複数存在する場合が多いこと，見出しは人間にとって目立つことの二点に注目する．本論文では次に，抽出された見出し構造の **Web** 検索への応用手法を提案する．提案手法は **Web** ページをその中で最高スコアのブロックのみに基づきランキングし，ブロックのスコアリングには祖先ブロックの見出しをメタデータとして考慮する．本論文ではその他の応用についても論じる．一つはクエリに対してその意図を特化・明確化するクエリのランキングを返すサブトピックランキングへの応用．一つは **Web** ページ中のキーワードの出現間の論理的関係の強さを測りページのスコアリングのために考慮する近接検索への応用．最後はユーザが検索結果ページ本文を読むか否かの判断に使用するページの簡潔な要約，いわゆるスニペットの生成への応用である．

This is a summary of "Web Search Based on Hierarchical Heading-Block Structure Analysis", which is Tomohiro Manabe's doctoral dissertation, in Japanese.

1 はじめに

Web の急速な発展の一因に，その自律分散性がある．自律分散性により，誰もが自由に Web ページを記述し，情報を発信できる一方，ページの記述は自然言語により，構造化の方法も多様で，その自動利用は難しい．したがって，Web 情報の自動利用は研究対象とされてきた．自動利用のためには，Web ページやページ群に共通する構造が重要である．リンク構造の有名な応用には PageRank や HITS があり，文構造は自然言語処理の分野で広く研究されている．一方，本研究が対象とするのは論理構造である．

本研究における論理構造とは，Web ページ中の文構造より大きな構造の総称である．論理構造には例えば，ヘッダ，メニュー，本文などを二次元的に配置するためのレイアウト構造が含まれる．本研究は見出し構造に注目する．これは，見出しのついたブロックの包含関係からなる論理構造である．見出しは Web ページ中のある一部の主題を表し，他の箇所とは容易に区別可能な箇所であり，ブロックとは見出しのついたページの一部である．ブロックは他のブロックを包含することがあるため，見出し構造は階層的になる．例として，典型的な階層的見出し構造をもつページを図 1 に示し，その階層的見出し構造を図 2 に示す．見出し構造は，レイアウト構造におけるメニューや本文をより細かく構造化し，文構造をもつ文を複数まとめてより粗く構造化する．

♡ 正会員　ヤフー株式会社　manabe@dl.kuis.kyoto-u.ac.jp

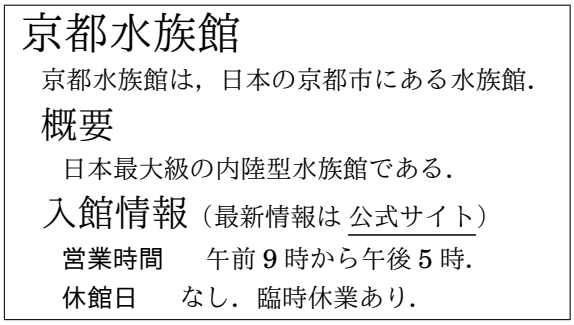

図 1: 階層的見出し構造をもつページの例．

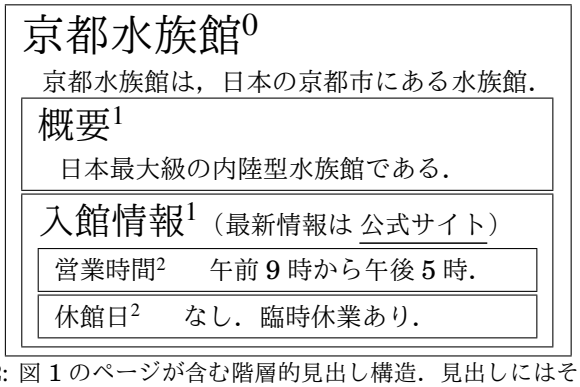

図 2: 図 1 のページが含む階層的見出し構造．見出しにはそのレベルを示す添字をつけ，ブロックは囲みによりそれぞれ表した．

Web において見出し構造はありふれたものである．最も普及した Web ページ記述言語である HTML に見出しを表すタグが定義されていることも，見出し構造の一般性を示している．実際，2 章で後述する通り，われわれの実験によれば，78%ものページが見出し構造をもつ．このように多くの Web ページが見出し構造をもつことから，その包括的な意味理解のためにも，意味理解を必要とする各種の自動利用のためにも，見出し構造の理解が必要である．しかし，見出し構造に関する既存研究は少ない．本研究では，これは，見出し構造の抽出が，実は研究対象となりうるほど難しいにもかかわらず，易しく思えることから，これまであまり論じられてこなかったためであると考える．したがって，本研究ではまず，Web ページ中の階層的見出し構造の抽出について論じる．

1.1 本稿の構成

本稿の構成は以下の通りである．まず次章では，Web ページ中の階層的見出し構造の抽出について述べる．これには，本研究全体の基盤となる見出しに基づくページセグメンテーション (Heading-based page segmentation, HEPS) 手法の概要が含まれる．本稿では続いて，HEPS により抽出された階層的見出し構造の四つの応用手法について述べる．第 3 章では，ブロックをページのようにスコアリングし，その際ブロックの階層的見出しをもメタデータとして考慮することで，クエリに基づく Web ページランキングを改善する手法について述べる．第 4 章では，ブロックの階層的見出しがその主題を簡潔に要約することと，ブロックの記述量がその主題の重要度を示すことに着目し，サブトピックランキングを生成する手法について述べる．サブトピックランキングとは，クエリに対してその意図を特化・明確化する別のクエリのランキングである．第 5 章では，見出し中の語と対応するブロック中の語の間や，異なるブロック中の語の間の論理的距離は，単純な語数とは異なることを考慮し，近接検索を改善する手法について述べる．第 6 章では，ブロックが含む各文をスコアリングし，その際階層的見出しを考慮することで，クエリに基づく Web ページの検索結果スニペットを生成する手法について述べる．

2 Web ページ中の階層的見出し構造の抽出

ほとんどの Web ページは，タグによる DOM 木構造を持ち，その構造は自明に抽出可能である．しかし一般に，DOM 木構造は階層的見出し構造とは対応しない．まず，このことを示すため，われわれは，TREC 2009–2012 Web track [3] データセット中でいずれかのクエリに適合する文書のうち，ランダムに 1,321 ページを抽出し，7 名の被験者によって本文中の階層的見出し構造を抽出した．結果として，15,560 の見出しブロック対を得，そのうち 68%は見出しを表すタグによって囲まれてはいなかった．また，見出しを表すタグにより囲まれた箇所のうち，37%は見出しではなかった．このように，DOM 木構造は階層的見出し構造とは対応しないことが示され，その抽出の困難さが示された．また，78%のページが（ページ全体を除き）少なくとも一つのブロックを，35%のページが少なくとも一対の親子関係にあるブロック対を含んでいた．このように，階層的見出し構造の一般性も示された．

2.1 アイデア

われわれは，階層的見出し構造の抽出のため，ページ中には同じ見た目の見出しが複数存在する場合が多いこと，見出しは人間にとって目立つことの二点に注目した．見出しを抽出することができれば，対応するブロックとしては，その見出しの直前から，次の同等以上のレベルの見出しの直前までを抽出すればよい．

2.2 手法：Heading-based Page Segmentation

以上のアイデアに基づき，われわれは，次の三つの主要ステップから成る階層的見出し構造の抽出手法，HEPS を提案する [5].

第一に，HEPS は，入力の DOM 木構造中のテキスト・画像ノードを見た目により分別してノード集合を得る．この分別のためには，DOM 木構造上でルートから注目するノードまでのタグ名を書き並べたタグパス，注目するノードの CSS スタイル情報（文字サイズ，文字色など），そして画像ノードの高さを用いる．

第二に，HEPS は，ノード集合を目立つ順に整列する．この整列のためには，ノード間の包含関係に基づく制約を考慮するほか，ノード集合の要素の文字サイズ，文字の太さ，ノード集合中で最初に出現する要素の文書順をこの順で優先して利用する．

第三に，HEPS は，ノード集合を目立つ順に走査し，見出しの集合をヒューリスティックに判定し，判定された見出しの集合に対応するブロック集合を抽出する．このようなノード集合単位の判定により，HEPS は，ノード単位の判定を行う既存手法 [19] や，ノード対を比較していずれかが見出しであるかを判定する既存手法 [12] よりも多くの情報を考慮できる．ヒューリスティクスの一つは，見出しの内容はその対応するブロックを包含する上位のブロックごとにユニークであるというものである．このように見出しの集合の判定のために上位のブロックを利用するため，HEPS は，ノード集合を目立つ順（推測されるレベルの降順）に走査し，見出しの集合の発見ごとに対応するブロックを抽出する．

2.3 評価

前述のデータセットを用いて評価を行った．抽出された見出しとブロックそれぞれの適合率，再現率，F 値を尺度とした．ベースラインとしては，決定木学習による見出し抽出 [10]，タグ名に基づく見出し抽出，VIPS による階層的ブロック抽出 [1] を用いた．

見出しの抽出について，学習による既存手法は，学習データにないフォーマットの Web ページに対してはうまく働かなかった（F 値 0.15）．それに対して提案手法は，タグ名により明示された見出しの抽出と同等の精度（適合率 0.67 に対して 0.64）で，はるかに多くの真の見出し（再現率 0.32 に対して 0.57）を抽出し，総合的な性能でも上回った（F 値 0.43 に対して 0.60）．

ブロックの抽出について，VIPS は，われわれの定義するブロックをうまく抽出しなかった（F 値 0.11）．これに対して HEPS は，見出しの抽出と同等の性能（適合率 0.59，再現率 0.56，F 値 0.57）でブロックを抽出した．特に，見出しを正しく抽出した場合，その対応するブロックも高い精度で抽出した（適合率 0.77）．

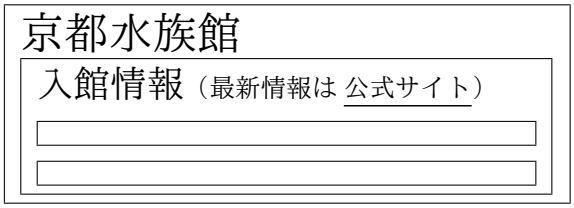

図 3: 図 1 中の入館情報ブロックについて，提案手法でブール検索を行う場合，階層的見出しが補われ，子ブロックが除かれる．

3 見出しを考慮したブロックベース Web 検索

Web ページ検索は，最も広く使用されている Web 自動利用システムの一つである．われわれは，前述の HEPS の第一の応用として，ブロックベース Web ページ検索手法を提案する．

3.1 アイデア

定義の通り，ブロックは，ある主題について書かれた情報の単位であり，検索の単位としても有用であると考えられる．通常の Web ページ検索手法が Web ページ単位のブール検索とスコアリングを行うのに対して，提案手法はブロック単位で同様の処理を行い，包含するブロックの最高スコアによってページをランキングする．ただし，ページをランキングするのは，単にのちにページランキングに関する標準的なデータセットを用いて手法を評価するためであり，提案手法は，ブロックのランキングを返し，ユーザが読むべきブロックまでも特定するシステムに直接応用可能である．

3.2 手法

提案手法は，ブロック単位の AND 検索を行う．すなわち，全ての検索語が出現するブロックのみを検索結果として抽出する．ただしその際，あるブロックを階層的に含む全てのブロックの見出しが元のブロックの主題を表すことを考慮し，それらの見出し中の検索語の出現は，元のブロックに関する出現として数える．さらに，子ブロックを読み飛ばす場合を考慮し，子ブロック中の検索語の出現は，元のブロックに関する出現として数えない．例えば，図 1 中の「入館情報」ブロックについて提案手法を適用する場合，図 3 中のテキストのみが AND 検索の対象となる．

提案手法は引き続き，抽出されたブロックを BM25F 関数 [14] を用いてスコアリングする．BM25F は各検索単位について複数のフィールドを考慮し，フィールドごとに検索語の出現に異なる重みを掛ける．提案手法が考慮するフィールドは，ブロックを包含するページのタイトル・URL・被リンクのアンカーテキスト，ブロックの本文・自身の見出し・親ブロックの見出し・親以外の祖先ブロックの見出しである．なお BM25F は BM25[15] の発展形でもあるため，フィールドの長さを考慮して正規化されている．

3.3 パラメータ最適化

TREC 2009–2012 Web track[3] のデータセットから 100 クエリを使用し，貪欲法 [9] により各フィールドの重みや正規化の強度など BM25F のパラメータを最適化した．提案手法の実装には，Apache Solr および BM25F プラグイン [13] を用いた．結果として，タイトルなど Web ページのスコアリングに効果的であることが既知 [14] のフィールドよりも，祖先ブロックの見出しが有意に重くなり，重要なメタデータであることが示された．

3.4 評価

同データセットの残る 100 クエリについて，前述のパラメータの値によるランキングを生成した．ベースラインとしては，ブロックを考慮しない BM25F[13] によるランキングを生成し，比較対象の既存手法としては，VIPS[1] が抽出したブロックを BM25[15] でスコアリングする BlockRank (BR)[2] によるランキングを生成した． nDCG@20 による評価では，BM25F が 0.173，BR が 0.175 に対して提案手法が 0.218 と大幅に上回り，また対応のある t 検定（同じクエリに関する同評価尺度の値を対応させる）によればベースラインとの差は危険率 5% で統計的有意であった．

表 1: 図 1 中の階層的見出しとその対応するブロックの文字数.

京都水族館	87
京都水族館 概要	18
京都水族館 入館情報	44
京都水族館 入館情報 営業時間	15
京都水族館 入館情報 休館日	13

4 見出しに着目したサブトピックランキング

ユーザが入力したキーワードクエリには検索意図があり，それを特化または明確化する別のキーワードクエリがサブトピックである．例えば，クエリ「京都 観光地」に対して，クエリ「京都水族館」,「京都 観光地 春」 はどちらもサブトピックでありうる．サブトピックランキングとは，あるクエリを入力したユーザの検索意図が，あるサブトピックを含む確率により，サブトピックをソートすることである．サブトピックランキングは，クエリ推薦のほか，検索結果の多様化にも応用可能である．例えば，ランキング上位のサブトピックについて，各サブトピックと関連の強い文書をそれぞれ返すことで，多様な検索意図を満たせるシステムが考えられる．これらの有用性から，サブトピックランキングは，NTCIR INTENT/IMine task などで広く研究されている [16].

4.1 アイデア

われわれは，(1) サブトピックの情報源として，階層的見出しを用いる．また，(2) 階層的見出しのスコアとして，その対応するブロックの記述量を用いる．これらはそれぞれ，(1) 検索意図を簡潔に要約するサブトピックと，ブロックの主題を簡潔に要約する階層的見出しとが類似するためであり，(2) 著者は読者のために文書を記述するのだから，読者が興味をもつ確率が高い内容をより多く記述すると考えられるためである．例として，図 1 中の階層的見出しを抽出し，対応するブロックの文字数を付して表 1 に示す.

4.2 手法

提案手法は以下の三段階でサブトピックランキングを生成する [8]. 第一に，提案手法は，コーパス中の各ブロックの重みを計算する．この段階においては，重みとして文字数を用いるものだけでなく，文字数の対数，親ブロックのスコア，子孫ブロックの数を用いるものを含めて四通りの手法を提案する．第二に，提案手法は，コーパス中の各階層的見出しを，その対応する全てのブロックの重みの総和によりスコアリングする．この段階においては，ページ単位・サイト単位でそれぞれスコアを正規化するか否かにより四通りの手法を提案する．第三に，提案手法は，階層的見出しをランキングする．この段階においては，一度だけスコアを計算し，それによりソートするものと，階層的見出しを一つランキングに含めるごとに，その対応するブロックをコーパスから除きスコアを計算し直すものの二通りの手法を提案する．後者の手法は，サブトピックランキングの多様化を意図したものである．各段階は任意に組み合わせ可能であり，提案手法は合計 32 通りとなる.

4.3 評価

評価のためには，NTCIR INTENT-2 task[16] の英語データセットと評価尺度を用い，このデータセットに含まれるベースラインを提案手法でリランキングし評価した．コーパスとしては，TREC 2012 Web track[3] のベースラインランキング中の Web ページを使用した．評価の結果，D-nDCG（適合率の尺度）, I-rec（再現率の尺度）, D♯-nDCG（前二者の算術平均で，総合的な尺度）のいずれにおいても，提案手法はベースラインを上回った．また，前述の検定によれば，全ての尺度における差が統計的有意であった．各段階については，文字数の対数によるブロックへの重みづけが有用であることがわかった一方，スコアの正規化の有無による性能差はなく，また多様化手法は有効ではなかった.

5 見出しを考慮した半距離に基づく近接検索

高精度なキーワード検索のためのアプローチの一つに，近接検索がある．近接検索を行う検索システムは，キーワードの出現頻度に加えて，出現の間の近接性をも考慮する．多くの近接検索手法は距離，すなわちキーワードの出現の間の語の出現の数を，非近接性の尺度として利用する．具体的な近接検索手法として，キーワードの出現の間の最短距離を考慮する MinDist [18]，12 個の距離の関数を遺伝的プログラミングで組み合わせた P6 [4]，ページを重なりのないスパンに分割し，スパンの両端の距離に基づくスコアを語の出現頻度にかけ合わせる Span [17] などがある.

これらの近接検索手法は階層的見出し構造を考慮していないが，一般には，階層的見出し構造が語と語の近接性に影響を与えている場合がある．例えば，図 1 中のページにおいて,「京都水族館」の最後の出現から「休館日」の出現までの距離は,「概要」の出現までの距離より大きい．しかし，このページ全体の見出しが「京都水族館」であることから，論理的には，両近接性は等しいといえる．このように，見出し中の語の出現と対応するブロック中の語の出現の論理的な非近接性は，距離よりも小さいと考えられる．また，同じページの例で,「5 時」と「休館日」の出現は隣接しているが，論理的には両出現間の関係は弱く，非近接性は大きいといえる．このように，異なるブロック中の語の出現間の論理的な非近接性は，距離よりも大きいと考えられる.

われわれは，階層的見出し構造を考慮する非近接性の尺度，Heading-Aware Semi-Distance (HASD) と，既存手法に HASD を取り入れた三つの見出し考慮近接検索手法を提案する [7].

5.1 アイデア：Heading-Aware Semi-Distance

o_1, o_2 間の HASD は，距離 *dist* を用いて以下の式で定義される.

$$hasd(o_1,o_2) = \begin{cases} dist(o_1,o_2)\cdot a_{hc} + b_{hc} & \text{if } hc(o_1,o_2) = \text{true}; \\ dist(o_1,o_2)\cdot a_{db} + b_{db} & \text{if } db(o_1,o_2) = \text{true}; \\ dist(o_1,o_2) & \text{otherwise.} \end{cases}$$

ただしここで，$hc(o_1,o_2)$ は片方の出現がある見出しに含まれ，もう片方がその対応するブロックに含まれる場合のみ真となり，$db(o_1,o_2)$ は $hc(o_1,o_2)$ が偽で，かつ o_1 と o_2 が異なるブロック（包含関係にあるブロックを含む）に含まれる場合のみ真となる.

5.2 手法：HA-MinDist, HA-P6, HA-Span

提案手法 Heading-Aware MinDist, Heading-Aware P6, Heading-Aware Span は，それぞれ対応する既存手法 [4, 17, 18] 中の距離を，提案する HASD で置換することで得られる．特に，Span では，ページをスパンに分割するためにもキーワードの出現間の距離を用いるが，この距離も HASD で置換する.

5.3 パラメータ最適化

TREC Web track[3] 2013–2014 のデータセットから 50 クエリを使用し，貪欲法 [9] により HASD のパラメータなど提案手法のパラメータを最適化した．$hc(o_1,o_2)$ が真の場合，二提案手法においては常に $hasd < dist$ となったが，HA-Span では HASD と距離に大きな差はなかった．$db(o_1,o_2)$ が真の場合，全提案手法において常に $dist < hasd$ となり，異なるブロック中の語の出現間の非近接性は距離よりも大きく，HASD の値に近いことが示された.

5.4 評価

同データセットの残る 50 クエリについて，前述のパラメータの値に基づくランキングを生成した．ベースラインとしては同データセットに含まれるランキングを用いたほか，比較対象として既存手法 [4, 17, 18] によるランキングも生成した．ERR-IA@20, α-nDCG@20, NRBP, MAP-IA で評価したところ，提案手法と対応する既存手法の全ての対について，全ての尺度で提案手法が対等以上の評価であった．このように，HASD の有用性が示された．また前述の検定によれば，全手法のうち HA-Span のみがベースラインを全尺度で統計的有意に改善し，その有用性も示された.

京都水族館
…… 入館情報（最新情報は公式サイト）営業時間 午前 9 時から午後 5 時．休館日 ……

図 4: キーワード「営業時間」に関する図 1 のスニペットの例.

京都水族館 > 入館情報 > 営業時間
午前 9 時から午後 5 時.

図 5: 同条件で，階層的見出し構造を考慮したスニペットの例.

6 見出しを考慮した検索結果スニペット生成

検索結果スニペットは，検索結果ページに表示される文書の要約で，ユーザが各文書の本文を閲覧するか否かの判断に用いられる．図 4 にスニペットを含む検索結果アイテムの例を示す．スニペット生成のためには，文書中の文をスコアリングし，スコアの高い文を優先的に抽出するアプローチが広く用いられている．その際には，文書をスコアリングする場合と同様，クエリ中のキーワードが出現する文のスコアを高くするのが通常である．

第 3 章で述べた通り，ブロックをスコアリングするにはその階層的見出しを考慮する必要があり，ブロック中の文のスコアリングについても同様であると考えられる．見出しを考慮するスニペット生成手法として，見出しそのものが重要な文であるとする手法 [20] が存在する．しかし一般には，見出し自体は非常に簡潔な要約で，主にその対応するブロックの内容を理解するための文脈を提供する．そのため，より新しい既存手法として，見出し中の語（見出し語）は，その対応するブロック中の文にとっての重要語であり，見出し語を含む文が重要であるとする手法 [11] がある．

6.1 アイデア

これまでのアイデアをまとめると，通常のアイデアは「キーワードを含む文が重要である」というもの，既存のアイデアは「見出し語を含む文が重要である」というものであった．これらに対しわれわれが提案するアイデアは,「ある文について，その階層的見出しがキーワードを含むなら，その文は重要である」とする．

6.2 手法

われわれは，これら通常・既存・提案アイデアをすべて取り入れたスニペット生成手法を提案する [6]．提案手法は BM25F [13] により文をスコアリングする．その際フィールドとしては，文そのものと文を含むブロックの階層的見出しの二つを考慮する．また，キーワードのほか，見出し語についてもスコアの計算に含める．さらに，フィールドにより異なる重みを付与するのと同様，キーワードと見出し語にも異なる重みを付与する．出力の際には，各文にその階層的見出しを補って表示する．提案手法の出力例を図 5 に示す．重みの設定は以下のように行った．語の重みについては既存研究 [11] に従い，キーワードは見出し語の 3 倍，キーワードかつ見出し語は 4 倍とした．フィールドの重みについては，知見が存在しないため等しいものとした．ただし階層的見出しフィールド中の見出し語の出現については，自明であるため無視した．

6.3 評価

INEX Snippet Retrieval track [21] の評価手法と，TREC 2012 Web track [3] のデータセットを用いて提案手法を評価した．ベースラインとして通常のアイデアのみを用いるもの，比較対象として通常のアイデアに加え既存・提案アイデアのいずれか一方を用いるものも評価した．再現率（適合なページのうち，スニペットのみから正しく適合と判定された割合）とネガティブ再現率（不適なページのうち，スニペットのみから正しく不適と判定された割合）の幾何平均によれば，一般のクエリに対しては既存・提案アイデアのいずれも有効ではなかった．一方，single クエリ（検索意図が比較的明らかなクエリ）や，4 語以上のクエリに対しては，既存・提案アイデアを共に用いる提案手法が最も有効であった．

[文献]

[1] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. VIPS: a vision-based page segmentation algorithm. Technical Report MSR–TR–2003–79, Microsoft, 2003.

[2] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Block-based web search. In *SIGIR*, pages 456–463, 2004.

[3] C. L. A. Clarke, N. Craswell, and E. M. Voorhees. Overview of the TREC 2012 web track. In *TREC*, 2012.

[4] R. Cummins and C. O'Riordan. Learning in a pairwise term-term proximity framework for information retrieval. In *SIGIR*, pages 251–258, 2009.

[5] T. Manabe and K. Tajima. Extracting logical hierarchical structure of HTML documents based on headings. *VLDB*, 8(12):1606–1617, 2015.

[6] T. Manabe and K. Tajima. Heading-aware snippet generation for web search. In *AIRS*, pages 188–200, 2015.

[7] T. Manabe and K. Tajima. Heading-aware proximity measure and its application to web search. *DBSJ Journal*, 14(2):1–6, 2016.

[8] T. Manabe and K. Tajima. Subtopic ranking based on hierarchical headings. In *WEBIST*, pages 121–130, 2016.

[9] D. Metzler and W. Bruce Croft. Linear feature-based models for information retrieval. *Information Retrieval Journal*, 10(3):257–274, 2007.

[10] H. Okada and H. Arakawa. Automated extraction of non <h>-tagged headers in webpages by decision trees. In *SICE*, pages 2117–2120, 2011.

[11] F. C. Pembe and T. Güngör. Structure-preserving and query-biased document summarisation for web searching. *Online Inform. Rev.*, 33(4):696–719, 2009.

[12] F. C. Pembe and T. Güngör. A tree learning approach to web document sectional hierarchy extraction. In *ICAART*, pages 447–450, 2010.

[13] J. Pérez-Iglesias, J. R. Pérez-Agüera, V. Fresno, and Y. Z. Feinstein. Integrating the Probabilistic Models BM25/BM25F into Lucene. *CoRR*, abs/0911.5046, 2009.

[14] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *CIKM*, pages 42–49, 2004.

[15] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *SIGIR*, pages 232–241, 1994.

[16] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, and R. Song. Overview of the NTCIR-10 INTENT-2 task. In *NTCIR*, 2013.

[17] R. Song, M. J. Taylor, J.-R. Wen, H.-W. Hon, and Y. Yu. Viewing term proximity from a different perspective. In *ECIR*, pages 346–357, 2008.

[18] T. Tao and C. Zhai. An exploration of proximity measures in information retrieval. In *SIGIR*, pages 295–302, 2007.

[19] Y. Tatsumi and T. Asahi. Analyzing web page headings considering various presentation. In *WWW*, pages 956–957, 2005.

[20] A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *SIGIR*, pages 2–10, 1998.

[21] M. Trappett, S. Geva, A. Trotman, F. Scholer, and M. Sanderson. Overview of the INEX 2013 snippet retrieval track. In *CLEF*, 2013.

真鍋 知博 **Tomohiro MANABE**

2011 年 3 月，京都大学工学部情報学科卒業．2013 年 3 月，京都大学大学院情報学研究科社会情報学専攻修士課程修了．2016 年 3 月，同専攻博士後期課程修了．京都大学博士（情報学）．2016 年 4 月より現職．テキスト検索システムに関する研究，開発に従事．