（ 続紙 １ ）

京都大学	博士（情報学）	氏名	真鍋　知博
論文題目	Web Search Based on Hierarchical Heading-Block Structure Analysis （階層的な見出しブロック構造の分析に基づくWeb検索）		

（論文内容の要旨）

本論文では、文書集合から必要な文書を検索する際に、各文書内に現れる、見出し付きブロック群が成す階層構造の情報を用いて、検索精度の向上と検索作業の効率化を実現する技術について論じる。

本論文では様々な種類の文書の中から、現在、社会的に重要な情報媒体となっているインターネット上のＷｅｂ文書を取り上げ、Ｗｅｂ文書検索の精度向上と作業効率化を目標として手法の開発および評価を行う。ただし、本論文で注目する階層的見出しブロック構造は様々な文書において用いられる、最も重要な文書内論理構造の一つであり、本論文で提案する手法はいずれも多様な文書群に対して適用可能と期待される。

Ｗｅｂ文書においては、そこに含まれる階層的見出しブロック構造は必ずしも自明ではない。そこで本論文では、まず、Ｗｅｂ文書より階層的見出しブロック構造を自動的に抽出する手法について論じる。続いて、抽出された構造の情報を用いてＷｅｂ検索の精度向上と作業効率化を実現するための、四つの手法について論じる。一つ目は、ユーザの検索トピックを表す検索フレーズが与えられた際に、そのサブトピックとなるフレーズを推薦するサブトピック推薦に関する手法、二つ目は複数の検索語により表現される検索要求が与えられた際に、それら検索語の近接性を考慮して検索結果のランキングを行う近接検索に関する手法、三つ目は文書全体ではなくブロックを単位として検索要求との一致度を評価することにより検索結果のランキングを行うブロック検索に関する手法、四つ目は検索結果中の各文書の要約を生成してユーザの文書選択作業を支援するスニペット生成に関する手法である。

最初の課題である階層的見出しブロック構造の自動抽出に関しては、文書著者は多様な方法で見出しを記述するが、読者がその外観デザインに基づいて見出しを判別できるよう記述するはずであるという仮定のもと、外観デザインこそが見出し抽出において唯一、普遍的に信頼できる情報であると考え、外観デザインの情報を用いて見出しを抽出する手法を提案する。

また、本提案手法は、一つの文書内では同レベルの見出しは同じ外観デザインを用いて表現されるという仮定のもと、まず、見出しの候補を同じ外観デザインを持つ集合に分類し、この集合の単位で、それが見出しかどうかの判定を行う。集合単位で判定を行うことにより、候補単体ごとに判定する手法に比べ、大域的な階層構造との一貫性など多くの情報を利用できる。実験により、本提案手法が既存技術に比べて同等の精度を保ちつつ再現率を大幅に向上することを確認した。

次に、サブトピック推薦に関しては、文書中の階層的見出しの内容はサブトピックに対応するという仮定、および、著者が各見出しに対応するブロック内に記述する文書量は、その見出しが表すトピックの、読者にとっての重要性を反映するという仮定のもと、Ｗｅｂ文書群中に現れる階層的見出しと、それに対応するブロック内の記述量を用いて、主要なサブトピックとその重要度を推定する手法を提案する。実験によ

り、本提案手法は、商用検索エンジンが検索ログに基づいて生成するクエリ推薦をそのままサブトピック推薦とみなして利用する手法に比べ、サブトピックのランキング精度を改善することを確認した。

続いて、近接検索に関しては、見出しとその対応するブロック内に出現する二つの検索語の間、あるいは、無関係な二つのブロック内に出現する二つの検索語の間の論理的距離は、それぞれ、検索語間の単語数に基づく物理的距離の一次関数で近似できると仮定し、そのような論理的距離に基づく検索ランキング手法のパラメータを最適化することで、各一次関数がどのような関数になるかを調べた。その結果、見出しとその対応するブロック内に出現する検索語対の論理的距離はほぼ物理的距離に等しいこと、無関係な二つのブロック内に出現する検索語対の距離は、それらが常に無関係であるとみなしてよいほど遠いということを明らかにした。

続いて、ブロック検索に関しては、各ブロックを、その内部、自身の見出し、親の見出し、その他の祖先の見出し、文書全体のタイトルなどの要素からなるとみなし、検索ランキングを最適化する各要素の重みを求めることによって、自身や親、祖先の見出しはタイトル以上に重要であること、自身の見出し以上に親や祖先の見出が重要であることを明らかにした。

最後に、スニペット生成に関しては、ある検索に対するスニペットにおいては、その検索語が含まれる見出しの子孫ブロック中の文を含めることが有効であるという仮定に基づいたスニペット生成手法を提案し、実験の結果、本提案手法は、検索意図が非常に明確な検索や検索語数が４以上である検索に対しては、既存手法と比べスニペットの有用性を向上することを確認した。

注)論文内容の要旨と論文審査の結果の要旨は１頁を３８字×３６行で作成し、合わせて、３，０００字を標準とすること。
論文内容の要旨を英語で記入する場合は、４００～１，１００wordsで作成し
審査結果の要旨は日本語５００～２，０００字程度で作成すること。

（ 続紙 ２ ）

（論文審査の結果の要旨）

本論文はＷｅｂ文書内の見出し付きブロックが成す階層構造の情報を自動的に抽出し、この情報をＷｅｂ文書検索の精度向上および検索作業効率化のために用いる手法に関するものである。Ｗｅｂ文書は、現在、もっとも重要な情報源の一つとなっており、そこから必要な情報を入手するためのＷｅｂ検索タスクが日々大量に実行されている。そのため、これらのタスクの検索結果の精度向上や作業の効率化は社会全体の生産性の向上につながる。よって、本論文で論じられている課題はたいへん社会的な重要度の高い課題であるといえる。また、本論文が注目する階層的見出しブロック構造は、約８割のＷｅｂ文書に存在することが本論文で示されている。よって、この情報を用いて精度向上や作業効率化を実現する本論文の提案手法は、Ｗｅｂ検索に関する様々な手法の中でも特に適用できる検索タスクの範囲が広く、重要度の高いものであるといえる。また、階層的見出しブロック構造自体はＷｅｂ文書以外でも広く使われており、普遍性も高い課題であるといえる。

本論文では、まず、階層的見出しブロック構造の自動抽出手法を提案し、続いて、サブトピック推薦、近接検索、ブロック検索、スニペット生成という四つの課題について、抽出した構造の情報を有効利用する手法を提案している。提案手法は、いずれも既存手法に比べて、多くの場合あるいは特定の場合において、同等あるいは上回る性能を達成することが信頼性の高い実験によって示されている。

この中でも特に、階層的見出しブロック構造の自動抽出は、それが様々なＷｅｂ文書処理の際の前処理として重要で、その抽出精度がその後の処理の精度に大きく影響する処理であるにも関わらず、これまで研究が十分に行われていなかったものであり、本研究の貢献は大きい。また、ブロック検索に関する提案手法は実験でこれまでのＷｅｂ検索手法コンテストの首位の手法を上回る精度を達成しており、Ｗｅｂ検索手法の研究がこれまで長年なされてきて精度の改善が容易ではないことを考えれば、古いが依然として重要である問題に対しての特に重要な成果である。

また、本論文では、これらの手法の提案以外にも、実験の結果からいくつかの興味深い知見を得ており、例えば、著者による各サブトピックについての文書記述量は、読者にとってのそのサブトピックの重要性を一定程度反映していると考えられること、文書中の各ブロックのトピックの判定において、そのブロックの親ブロックや祖先ブロックの見出しは、そのブロック自身の見出し以上に重要であることなどが示されている。これらの知見も、学術的価値の高いものである。

また、本論文で示されている研究成果をまとめた論文として、これまでに四本の論文が査読付きの国際学術雑誌、または、国際学術会議等に採択されており、国際的な研究コミュニティにおいてもその価値が評価されていることが確認できる。

よって、本論文は博士（情報学）の学位論文として価値あるものと認める。

また、平成２６年２月２３日論文内容とその関連事項に関する口頭試問を行った結果合格と認めた。

注）論文審査の結果の要旨の結句には、学位論文の審査についての認定を明記すること。
更に、試問の結果の要旨（例えば「平成　年　月　日論文内容とそれに関連した口頭試問を行った結果合格と認めた。」）を付け加えること。

Webでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。
要旨公開可能日：　　年　　月　　日以降