様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１９　日現在

機関番号：　１２６０８
研究種目：　若手研究(B)
研究期間：　2015～2017
課題番号：　１５Ｋ２０９９０

研究課題名（和文）ユーザの検索目的の推定と検索目的に基づく情報提示手法の提案

研究課題名（英文）Estimation of user's intent and presentation method of search results based on the intent

研究代表者

欅　惇志（Keyaki, Atsushi）

東京工業大学・情報理工学院・助教

研究者番号：００７３３９５８

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本課題では，タスク指向型情報検索システム実現に向けて，主にその要素技術の提案・改善に取り組んだ．これらの過程で，Web 文書にも適用可能な正確な部分文書検索技術，クエリキーワードに対する正確な品詞付与技術，異種情報源の正確な統合手法を提案した．また，タスクごとにその検索難易度や適切な検索結果の情報粒度の分析を行った．これらの研究の成果により，タスク指向型情報検索システムの実現に向けて大きく前進することができた．

研究成果の概要（英文）：In this project, we aimed at materializing a task-oriented information retrieval system. For the purpose, we proposed an accurate element-based retrieval method for Web documents, an accurate part-of-speech tagging method for Web queries, and an accurate information integration method for heterogeneous information sources. Additionally, we conducted per-task analysis and revealed difficulty and appropriate information granularity for search results of each task. As the results of this project, we made substantial progress to materialize a task-oriented information retrieval system.

研究分野：　情報検索

キーワード：　タスク指向型情報検索　部分文書検索　クエリの品詞解析　コピュラ　MobileClick

１．研究開始当初の背景

現在 Web 上には大量の文書が存在し，IDC 社の調査によると 2020 年の総デジタルデータの蓄積量は 40 ゼッタバイトに及ぶと報告されている．このような大規模データは個人で扱える情報の規模を遥かに上回っており，Web から情報を取得する際に情報検索システムを利用することは，現代社会における社会活動の一部といっても過言ではない．

検索システム利用者は情報要求を解決すべく，クエリを繰り返し修正しながら，タイトルや概要文を手掛かりに選択した文書から，求める情報を探し出す．ユーザはこの過程に大きな無駄を感じているという報告からも，ユーザがクエリを入力すれば，各文書全体を閲覧することなく，直接的に求める情報を得ることで，Web 検索システムの利便性は飛躍的に向上する．ユーザの欲する情報を直接的に，かつ，正確に提示するためには，文書全体から情報要求を満たす一部分を検索結果として提示する必要がある．このような経緯から，現在は文書単位よりも細かな検索粒度の情報検索が盛んに取り組まれている．代表的アプローチの一つは，適合コンテンツをより多く含み，非適合コンテンツを極力含まない部分（章，節，小節など，文書構造を表す任意の粒度）を抽出する（部分文書検索）ことである．その際，文書の構造情報や統計情報を利用し，単体で理解可能な文書中の一部分（部分文書）が提示される．

その一方で，Web 文書は多種多様な文書集合から構成され，Web 検索においては QA 質問検索，語の意味や定義検索，人名検索など，様々なタスクが存在する．クエリ意図に適切な検索結果を提示する上で，これらタスクを推定することは極めて重要である．実際，タスクごとに特化した方針に沿って検索結果を構築することで，検索精度が向上することが報告されている[1].

[1] I. Kang et al., “Query Type Classification for Web Document Retrieval”, in Proc. of the 26th SIGIR, pp.64–71, 2003.

２．研究の目的

Web 検索を行う際のユーザの多種多様な目的に沿い，適切な検索結果の提示を行うことを目的として，タスク指向型情報検索システムの実現に取り組む．従来の文書（ページ）を粒度とした情報検索ではなく，文書中から情報要求を満たすコンテンツのみを提示するため，ユーザの情報検索における労力を軽減し，検索時間を短縮する．

上記の Web 検索システム実現のため，まずは検索システム利用者が入力した検索質問（クエリ）を分析して，QA 質問検索や人名検索などといった，検索課題のタイプ（タスク）を推定する．

クエリのタスクが推定されれば，高精度な部分文書検索技術によって抽出された，単体で理解可能かつクエリに対する適合コンテンツに対して，タスクごとに適切な方針にて検索結果の構築を行うことで，ユーザの検索目的を反映させた検索システムを開発する．

３．研究の方法

前述の目的を達成するため，本課題では，各種要素技術として，(1) Web 文書にも適用可能な正確な部分文書検索技術，(2) クエリキーワードに対する正確な品詞付与技術，(3) 異種情報源の正確な統合手法を提案した．また，(4) タスクごとにその検索難易度や適切な検索結果の情報粒度の分析を行った．

(1) タスク指向型情報検索システム構築において，文書中からクエリに適合する箇所を正確に抽出する技術の実在が前提となる．申請者は日本学術振興会特別研究員としての採用期間に，部分文書検索のアプローチにて Wikipedia などの高度に構造化された Web 文書に対して当該課題に取り組んで世界最高水準の部分文書検索システムを提案した．本課題ではそれらの知見を有効活用し，Web 文書に特化したアプローチを提案した．すなわち，部分文書検索技術で上位にランキングされた部分文書と類似度の高い箇所を適合記述として選択するランキング手法や，より情報要求に合致する検索結果を提示するためのクエリ拡張手法の提案を行った．

(2) クエリのタスク推定を行う上で，クエリキーワードの品詞情報は重要な手がかり情報として多用される．しかしながら，クエリには，語順がランダムであること，また，大文字情報が欠如するという特徴に起因し，既存の品詞分析技術(形態素解析技術)では正確に品詞を付与することができない．従って，本課題では，大規模 Web コーパスに対して付与された品詞解析結果を用いて，Web クエリに対する高精度な品詞情報付与手法の提案を行った．自然文単位の品詞分析精度は実用的なレベルに達しているため，大規模 Web コーパスに付与される品詞分析結果は正確であることに着目して，大規模 Web コーパス中のクエリキーワードとその品詞の組合せが付与される確率を計測して，より尤度の高い品詞をクエリキーワードに付与する．

(3) ユーザの情報要求の多様化の結果，情報検索システムにおいては，さまざまな評価基準（語の重み付け手法やリンク分析手法，文書の新鮮度やユーザの

嗜好との一致度など）を包括的に考慮して，最終的な検索結果の提示を行う．これらの評価尺度の統合において，既存の線形結合では，複雑な依存関係を反映することができないという問題がある．そこで，金融工学におけるリスク分析において用いられるコピュラを情報検索に適応することで，複雑な依存関係を考慮した複数の評価尺度を考慮した情報検索システムの実現を目指す．その際，既存の手法では検討されていなかった，ユーザの情報要求の多様性・複雑性に着目し，ユーザの情報要求を満たす箇所が複数クラスタに分割されている場合にも適切に反映させることが可能なアルゴリズムの提案を行なった．

(4) 申請者は，タスクごとに適切な検索結果の提示形式が異なるという仮定のもと，タスクごとに適合記述の情報粒度の計測を行った．その際，多量なデータに対して分析を行うため，クラウドソーシングにより，各適合記述に対して人手で情報粒度（語句，複合語，フレーズ，文）の付与を依頼した．

４．研究成果

(1) Web 文書中から適合箇所を抽出し，二層構造に要約して提示するタスクである，NTCIR-12 の MobileClick-2 にて，世界各国の主要な研究機関からの参加者の中で，最高の精度を達成することができた．また，これらの成果は，英文論文誌として報告を行い，論文賞 runners-up として選出された．

(2) 最先端の既存手法である，短文に対しての分析に特化したモデルを用いた形態素解析ツールと比較して，提案手法は 8% 精度を向上させることに成功した．その際，既存技術において特に精度の低かった，一般名詞と固有名詞の品詞判別を適切に行うことができた．これらの成果は，国際会議である SAC 2017 に採択され（採択率 23.43%），報告を行った．

(3) Web 検索システムにおいて重要な評価尺度である，検索結果上位における検索結果において，提案手法は既存の手法と比較して 5% 以上高精度に検索可能であるという結果が得られた．これらの成果は，国際会議である DEXA 2016 に採択され（採択率 28%），報告を行った．

(4) クラウドソーシングによって適合記述に情報粒度の付与を行った結果，より難易度の高いタスクにおいてはより情報粒度が高くなるという結果が得られた．また，同様に難易度の高いタスクほど情報の構造化が困難であるという結果も得られた．

これらの研究の成果により，タスク指向型情報検索システムの実現に向けて大きく前進することができた．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

① Atsushi Keyaki, Jun Miyazaki, Kenji Hatano: “Effective Mobile Search using Element-based Retrieval”, 情報処理学会論文誌：データベース, Vol.10, No.3 (TOD75), 2017. 査読あり

〔学会発表〕（計 33 件）

① Toshiaki Wakatsuki, Atsushi Keyaki, and Jun Miyazaki: “A Case for Term Weighting using a Dictionary on GPUs”, proceedings of the 28th International Conference on Database and Expert Systems Applications (DEXA 2017), France, August, 2017. 査読あり

② Shuhei Kishida, Seiji Ueda, Atsushi Keyaki, and Jun Miyazaki: “Skyline-based Recommendation Considering User Preferences”, proceedings of the Asia Pacific Web and Web-Age Information Management Joint Conference on Web and Big Data (APWeb-WAIM 2017), China, July, 2017. 査読あり

③ Atsushi Keyaki and Jun Miyazaki: “Part-of-speech Tagging for Web Search Queries Using a Large-scale Web Corpus”, proceedings of the 32nd ACM Symposium on Applied Computing (SAC2017), pp.931-937, Morocco, April, 2017. 査読あり

④ Yume Sasaki, Takuya Komatsuda, Atsushi Keyaki, and Jun Miyazaki: “A New Readability Measure for Web Documents and its Evaluation on an Effective Web Search Engine”, proceedings of the 18th International Conference on Information Integration and Web-based Applications & Services (iiWAS2016), pp.357-364, Singapore,

November, 2016. 査読あり

⑤ Takuya Komatsuda, Atsushi Keyaki, and Jun Miyazaki: “A Score Fusion Method Using a Mixture Copula”, 27th International Conference on Database and Expert Systems Applications (DEXA 2016), Volume 9828 of LNCS, pp.216-232, Porto, September 2016. 査読あり

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
欅　惇志（KEYAKI, Atsushi）
東京工業大学・情報理工学院・助教
研究者番号：00733958

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）