様式 C-19

科学研究費補助金研究成果報告書

平成２３年６月３日現在

機関番号：１２６１２
研究種目：基盤研究（C）
研究期間：2008～2010
課題番号：20500086
研究課題名（和文）　検索システムにおける個別 WEB 応用への対応化技術の研究

研究課題名（英文）　Web Application Technology in Retrieval System

研究代表者 尾内　理紀夫（ONAI RIKIO）
電気通信大学・大学院情報理工学研究科・教授
研究者番号：70323871

研究成果の概要（和文）：
研究の結果、システム全体は、文書収集・登録部(新規収集クローラと更新クローラ、文書登録モジュール)と検索部(検索バックエンド、インデクサ、スコア作成モジュール)から構成されることとなった。２種類のクローラで採用するオープンソースソフトウェアについて検討し、新規収集クローラはHeritrixを使用することし、更新クローラ、文書登録モジュールを実装した。さらに規模の拡大における負荷軽減、スケーラビリティ、耐障害性について検討し、Hadoopを導入し、HDFSで管理するようにした。MapReduceによるインデキシングの高速化を図り、従来に比較し、インデックスサイズはほぼ同等で約15倍の速度向上を実現した。以上、成果として本方式の有効性を検証した。

研究成果の概要（英文）：
Results of the study, the entire system was composed from a document collection・registration unit s (newly acquired crawler, updated crawler, and a document registration module) and a search unit (back-end search, indexer, and scoring module).We consider adopting open source software in the two crawlers, Heritrix crawler was adopted as newly acquired crawler, and the updated crawler and the document registration module were implemented. Load reduction, scalability, and fault tolerance examined, Hadoop and HDFS were introduced. Aims to speed up the indexing with MapReduce, compared to our conventional method, improved to about 15 times faster (size of the index was the same). The validity of this method was gotten as a result.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,200,000	360,000	1,560,000
2009 年度	1,300,000	390,000	1,690,000
2010 年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：総合領域
科研費の分科・細目：情報学・メディア情報学　データベース
キーワード：情報検索

１．　研究開始当初の背景

研究開始当初、日本語対象のオープンソースの全文検索システムの研究開発やパーソナル化を目指した検索システムの研究開発は活発に行なわれていた。これはユーザが検索対象サイトを登録し、検索対象を絞り込み、高精度の検索を可能にするという方式である。例えば、ROLLYOやGoogle Custom Search Engineは、ユーザが登録した複数のサイトやページに絞り込んで検索を行う、パーソナルな検索エンジンであり、このパーソナルな検索エンジンを共有することも可能である。これらはクローラ部とインデックス部をパーソナル化している。しかし個々のWEB応用の特性に対応した個別の検索対応化(一種のパーソナル化)を試みた検索システムはなく、検索部をパーソナル化し、スコアリングを変更してもインデックスの再構築を不要としているものもない。対象はテキストのみであり、画像検索WEB応用の特性を考慮し、検索部をそれに対応化したものもない。

評判情報検索システムにおいて、多くは領域依存型の検索手法であった(立石ら，”インターネットからの評判情報検索,”人工知能学会誌, Vol.19, No.3, pp.317-322, 2004.、K.Daveら,”Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” Proc. 12th Int. World Wide Web Conf.,pp.519-528, 2003.)。これらは領域知識として予め評価表現辞書を作成した分野に対する検索のみのため，汎用性において劣り、個人による評判情報とプロライター等による評判情報の区別が困難である。一方、我々の手法は個人記述のテキストにおける典型的な特徴を利用するため、領域独立型であり、個人による評判情報のみを効率的に収集可能である。

さらに、Why型質問を対象としたシステムはほとんどない．TRECに出場している質問応答システムでも、Why型質問に対応したアルゴリズムを実装しているシステムはない。わずかにSrihariらのシステムは、質問に“why”もしくは“for what”がある場合、Why型質問とみなし回答抽出を試みる(R.Srihari ら :Information extraction supported question answering, Proceedings of the 8th Text Retrieval Conference (TREC-8), pp.185-196, 1999)。しかし我々の手法のように回答位置特定アルゴリズムを明確にし、回答位置の特定は行っていない。日本では検索エンジンgooが日本語自然文検索実験サービスWebAnswersを公開していたが、技術の詳細は不明であり、実験は2005年5月に終了し、研究開始当初においては使用することができなかった。

blogを対象とした研究としてblogWatcherがあるが、我々の手法とは異なり，新たな興味を喚起する記事の推薦はしない．blog記事推薦サービスとしてBulkfeeds、News & Blog Search、So-net blogなどがあるが、類似記事の推薦であり、我々の手法のように意外性といった尺度での記事推薦は行っていない。

次に、背景を踏まえ、本研究の着想について述べる。

blogに投稿される画像情報に着目し、画像情報を含むblog記事を収集・解析し、それらを検索するシステム「もぶろげっと」を開発し、発表し、サムネイル画像取得機能をもぶろげっとAPIとして無償公開した。もぶろげっとでは、全文検索システムにNamazuを使用したが、インデクシング速度、検索速度とも不十分であるという知見を得た。

また、ユーザのblog記事を元に、そのユーザの新たな興味に繋がる他人のblog記事を推薦するシステムMineBlogを開発し、発表していた。関連(類似)性、意外(相違)性、話題性の3つの尺度を導入し、これらによりblog記事をスコアリングした。実験により、推薦記事の約2つに1つはユー ザの新たな興味に繋がる推薦記事であるという結果を得た。MineBlogではクローラ部に10種類のblogサービスを登録し、それに絞り込んでクローリングを行なった。しかし、クローラ部は汎用とし、検索部にblog記事判定機能を導入する方式のほうが精度向上が期待できることが判明した。

さらに、Why型質問に対する回答の自動抽出を目指し、回答文の出現位置をWEBページ

から特定する回答位置特定アルゴリズムを提案し、システムRE:Whyを実装、評価し、発表していた。評価実験の結果、一定程度の適合率、再現率を得ることができたが、RE:WhyはGoogle APIを使用しているため、それ特有のスコアリングが精度向上の制約となることも判明した。検索システムを開発し、その検索部に、回答位置の特定を支援する特徴語とその重みによるスコアリング機能を導入することにより、回答文を含むWEBページが上位にランキング可能になることが予想された。

評判情報を検索するために、WEBページから個人ページを自動抽出する手法を開発し、既発表の評価実験によりキーワード型検索エンジンを単独で使用する場合と比較して平均2.1 倍の精度向上が確認できた。しかし、Google APIを使用しているため、それ特有のスコアリングが精度向上の制約となることも判明した。検索システムを開発し、その検索部に、個人記述に特有な特徴語とその重みによるスコアリング機能を導入することにより、上位に評判情報記事がランキング可能になることが予想された。

これらの我々の研究成果を踏まえ、負荷軽減、スケーラビリティ、耐障害性の各機能を具備し、規模の拡張に対応した検索エンジンの研究を着想した。

２．研究の目的

本研究の申請時においても日常生活においてWEB検索は必須のものであり、活発に研究が展開されていた。研究代表者と研究分担者が主催する研究室では、WEB応用の研究を展開しており、そこにおいて、オープンソースの全文検索システム、Google APIなどを利用してきた。しかしAPIには様々な制約条件があり、十分な研究の展開にとって障害となっていた。また、オープンソースの全文検索システムも性能に不十分な点があった。そこで、下記の目的の研究を構想した。

個々のWEB応用の特性に対応した検索対応化機能を具備した検索システムをオープンソースベースに検討、実装する。対応化は検索部のスコアリングの対応化により実現する方式を検討する。申請者らによって開発されたWEB応用システムをテストベッドとして、有効性を検証する。

３．研究の方法

初年度である20年度は開発方針、システムの全体構成、検索部の基本部分などについて検討し、一部実装を行う。システムに関してはデータ量の増大に対応するため、データベースやインデックスを分割し、複数の計算機に分散して保存する方針が適切かどうかを検討する。また、開発負担をできるだけ抑えるため、様々なオープンソースソフトウェアを組み合わせる枠組みを検討する。そのため,オープンソースソフトウェア自身のソースコードを直接変更しないでシステム構築をすることが可能かどうかを検討する。特に、データ量の増大に対処するために必要となる機能のうち、データベースの分割とインデックスの分割の機能は単純にオープンソースソフトウェアを組み合わせるだけで実現できるかどうかを検討する。

21年度は、20年度の研究成果をベースに、システムをプロトタイピングし、動作などを検証する。特に、収集したページをインデックスに登録するインデクサなどをオープンソースソフトウェアで構成することの詳細検討を行う。具体的には、Heritrix、Apache Lucene,MySQLなどのオープンソースソフトウェアをベースにした検索エンジンの検討を行う。さらに、収集した各ページに対するスコアを作成するスコア作成モジュールの内部の検討を行う。また、インデックスから検索を行い、スコアを読み込んで検索結果のソートを行う検索バックエンド内部の検討を進める。

22年度は、20年度と21年度の成果をベースに、規模の拡大における負荷軽減、スケーラビリティ、耐障害性について検討する。また、WEBページのデータ管理について検討する。さらに、MapReduceによるインデキシングの高速化について検討し、実装、評価する。

４．研究成果

20年度は開発方針、システムの全体構成、検索部の基本部分などについて検討し、一部実装を開始した。まず、本システムではデータ量の増大に対応するため、データベースや

インデックスを分割し、複数の計算機に分散して保存する方針を採用することとした。また、開発負担をできるだけ抑えるため、様々なオープンソースソフトウェアを組み合わせる枠組みを採用することとした。そのため，オープンソースソフトウェア自身のソースコードを直接変更することはしないこととした。ただし、データ量の増大に対処するために必要となる機能のうち、データベースの分割とインデックスの分割の機能は単純にオープンソースソフトウェアを組み合わせるだけでは実現できないため、新たにモジュールを作成し、これら作成したモジュールから既存のオープンソースソフトウェアを呼び出すことで機能を実現することとした。

システム全体構成は，大きく分けて文書収集・登録部と検索部から構成される。文書収集・登録部は新規収集クローラと更新クローラ、文書登録モジュールに分かれており、検索部はインデクサ、スコア作成モジュール、検索バックエンドに分かれている。20年度は特に検索部の基本検討を行った。

21年度は、20年度の研究成果であるシステム全体構成をベースに、各部の検討を進め、一部を実装し、動作などを検証した。また、収集したページをインデックスに登録するインデクサなどをオープンソースソフトウェアで構成するための検討を行った。具体的には、Heritrix、Apache Lucene，MySQLなどのオープンソースソフトウェアをベースにした検索エンジンの検討を行った。収集した各ページに対するスコアを作成するスコア作成モジュールの内部の検討を行い、Luceneが標準で提供しているスコアリングをそのまま使用する方式と、利用者が独自で定義できるスコアリングを使用する方式を選択できるようにした。

インデックスから検索を行い、スコアを読み込んで検索結果のソートを行う検索バックエンドの内部検討を進め、LuceneとSenを使用することとした。20年度における研究により、文書収集・登録部は、新規収集クローラと更新クローラ、文書登録モジュールから構成されることとなったが、21年度は、これら二種類のクローラで採用するオープンソースソフトウェアなどについて検討し、新規収集クローラはHeritrixを使用することとした。更新クローラについては実装を進めた。また文書登録モジュールを実装した。

22年度は、20年度と21年度の成果をベースに、規模の拡大における負荷軽減、スケーラビリティ、耐障害性について検討し、Hadoopを導入した。21年度までは、WEBページのデータはMySQLを使用してデータベースで管理していたが、22年度においては、HDFSで管理するようにした。具体的には、クローリングしたWEBページ情報をHDFS上のHBaseで管理する。HBaseはHDFS上で動作するためWEBテーブルのデータは自動分割され、クラスタ内に分散して保持され、データ量の増加に対するスケーラビリティが得られた。またHDFS上にデータが保持されるため障害に対して安全である。さらにMapReduceによるインデキシングの高速化を図り、21年度に比較し、約15倍の速度向上を実現した。この時のインデックスサイズはほぼ同等であった。

以上、結果として本方式の有効性を検証した。

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）

早坂良太，林貴宏，尾内理紀夫「規模の拡張に対応した検索エンジンの開発」日本ソフトウェア科学会、コンピュータソフトウェア，査読有、Vol.26，No.4，pp.138-156，2009年

６．研究組織

(1)研究代表者

尾内　理紀夫（ONAI RIKIO）
電気通信大学・大学院情報理工学研究科・教授
研究者番号：70323871

(2)研究分担者

林　貴宏　（HAYASHI TAKAHIRO）
新潟大学・自然科学系・准教授
研究者番号：60342490