人とコンピュータによる「協調型情報検索」のための特徴抽出手法に関する研究

大野，麻子

(Degree)
博士（学術）

(Date of Degree)
2009-03-25

(Date of Publication)
2011-11-11

(Resource Type)
doctoral thesis

(Report Number)
甲4619

(URL)
https://hdl.handle.net/20.500.14094/D1004619

氏　　名	大野　麻子
博士の専攻分野の名称	博士（学術）
学 位 記 番 号	博い第 4619 号
学位授与の 要 件	学位規則第 5 条第 1 項該当
学位授与の 日 付	平成 21 年 3 月 25 日

【 学位論文題目 】

人とコンピュータによる「協調型情報検索」実現のための特徴抽出手法に関する研究

審 査 委 員

主 査　准教授　村尾　元
教　授　森下　淳也
教　授　鏑木　誠
教　授　大月　一弘
准教授　清光　英成

［課程博士用］

論文審査の結果の要旨

氏　　名	大野　麻子		
論文題目	人とコンピュータによる「協調型情報検索」実現のための 特徴抽出手法に関する研究		
判　　定	合　格（○）・不　合　格		
審査委員	区　分	職　名	氏　　名
	主　査	准教授	村尾　元
	副　査	教　授	大月　一弘
	副　査	教　授	鏑木　誠
	副　査	教　授	森下　淳也
	副　査	准教授	清光　英成

要　　旨

本論文はコンピュータ上のデータの特徴を数値化するための新しい手法を提案するものである．得られた数値を特徴量と呼ぶ．異なるデータから得られた特徴量を比較することにより，元のデータが互いに類似しているかどうかを判定できる．これを利用すると，コンピュータが持っている大量のデータから，利用者によって与えられたデータに類似したデータを見つけ出すことができる．これがコンピュータによる情報検索の仕組みである．

本論文が対象とする課題は，データのどのような特徴を数値化するのか，ということである．一般に，単一のデータには非常に多くの特徴がある．情報検索によって特定のデータ群を得ようと試みるとき，利用者によってコンピュータに与えられるデータと，検索対象データ群の特徴量が近くなければ，利用者の意図したような検索結果を得ることはできない．従って，どのような特徴を数値化し，特徴量とするかは，情報検索にとって非常に重要である．

本論文はこの課題の解決を利用者に委ねる．すなわち，情報検索に利用する特徴を利用者自身が選択できる新しい手法を提案する．本論文では，利用者自身が特徴を選択できる手法を用いた情報検索を「協調型情報検索」と呼んでいる．論文では２つの手法が提案されているが，いずれも検索したいデータのサンプルを利用者が与えることにより特徴を選択する手法である．

１つは，与えられたサンプル群とデータとの「距離」を定義し，これを特徴量として利用する手法であり，本論文では「FRefアルゴリズム」と呼んでいる．この手法では，利用者はサンプルを変更することで意図した検索結果が得られるように特徴量を調整することができる．

もう１つは，時系列的なデータを対象として，与えられたサンプル群の表面的な特徴の時間的な変化を確率モデルに学習させる手法であり，本論文では「CMアルゴリズム」と呼んでいる．この手法は，表面的な，つまり，利用者の目に触れる特徴を用いることが可能であり，これにより，利用者の意図した検索結果を得ようとしている．

本論文は６章から構成されている．第１章は本論文の目的と論文の構成について書かれている．第２章は研究の背景である．ここでは，対象としている課題を思いつくに至った経緯や，関連する研究の現在の状況と提案手法の位置づけについて，十分な量の既発行の文献を参照しながら説明されている．第３章以降は本研究で考案した新しい手法について説明が行われている．

まず，第３章では提案する２つの手法について説明されている．いずれの手法に対しても，数学的な表記も利用しながら，簡潔かつ十分な説明が与えられている．対象となるデータに依存しない一般的な説明が行われており，本論文を閲覧した研究者による再利用も可能である．

第４章，第５章は２つの手法それぞれを実際のデータに適用した実験について書かれている．実際のデータとして，工学部のプログラミングの授業で課題として提出されたプログラム及びデジタルカメラで無作為に撮影した写真が用いられている．これらに対して提案手法を適用し，情報検索を行うに必要最低限の性能を持っていること，利用者の意図する特徴を選択する能力を有することが示されている．実験方法および結果は正確かつ分かりやすくまとめられており，これらに基づいた考察は論理的である．

第６章は結論である．

論文全体を通して，利用者自身が特徴を選択できるという手法の新規性（第１章，第２章），その構成（第３章）と，その有用性（第４章，第５章）が論理的かつ明確に説明されている．

本論文の内容の一部は２件の国際論文誌を含む３件の査読付き論文誌に掲載されている．また，４件の査読付き国際会議での口頭発表が行われており，国内会議２件をあわせ，合計９件の発表が行われている．そのうち，１件の国際会議ではStudent Best Paper Awardを受賞，１件の国内会議では学生奨励賞を受賞している．１件は教育工学に関する学術会議であり，残る８件は情報分野に関する学術会議である．

本論文は，コンピュータによる情報検索に，利用者の判断を活用しようとする試みの礎的な研究について著述されたものである．本研究は，情報分野において学術的に価値あるのみならず，人の感性をコンピューティングに利用しようという学際的な学術分野に少なからぬ貢献を行うものである．また，第４章，第５章の実験では，本研究が教育分野においても有用であることが示されており，当該分野における貢献も期待される．

以上のような理由から，本審査委員会は，大野麻子氏に博士（学術）の学位を得る資格があると判定する．

論 文 内 容 の 要 旨

氏　　名　大野　麻子

指導教員氏名　村尾　元 准教授

論文題目（外国語の場合は，その和訳を併記すること。）

人とコンピュータによる「協調型情報検索」実現のための特徴抽出手法に関する研究

論文要旨

本研究では，人がコンピュータに対して自分の意図する通りの要求を専門知識無く伝えられるようなしくみをソフトウェア的に実現することを目指している．そのインターフェースとして，多くの人々が接点を持ち，用途も多岐に渡る情報検索を選び，研究対象とした．本研究において追求するものは，人とコンピュータによる「協調型情報検索」である．本研究では情報検索を「人とコンピュータのコミュニケーション・インタフェース」として位置づけ，ユーザの意図や好みに合った情報検索を難解な操作や知識を要求することなく実現することを目指している．このような情報検索の実現のためには，次の 2 つが必要であると考えられる．

i)人の意図を簡易な操作でシステムに伝えるための UI（ユーザ・インタフェース）

ii)人の意図に基づく特徴抽出，柔軟な特徴抽出

本研究ではこのうち ii)にあげられる特徴抽出手法として，1)参照を用いた特徴抽出手法，および 2)記述スタイルモデルに基づく特徴抽出手法の 2 つを開発した．

参照を用いた特徴抽出手法は，対象となるデータから直接特徴抽出を行うのではなく，複数の参照データと呼ばれる第三者のデータを用いて対象データの特徴を表現する．本手法は対象データに対し詳細な解析に基づく特徴抽出を行わず，参照データとの類似度に基づく相対的な特徴表現を採用しているため，一次元データとみなせるどのようなデータにも容易に実装可能である．データの特徴は参照データとの類似に基づいた相対的な尺度により表されるため，ユーザはその検索意図に応じて自由に類似性尺度を変更することができる．このため，簡易な操作によりユーザが意図する様々な特徴に着目した類似性検出を単独で実現することが可能である．本稿ではこれをもとにソースコード間類似性検出手法を開発した．特徴ベクトル算出までのプロセスと類似度の計算は完全に分離されているため，ソースコードの長さによらず高速な類似ソースコード検索を行うことができる．個人の小規模なリポジトリから，大規模なソフトウェアリポジトリまで，幅広い適用が期待できる．また，このような特徴をもつ本手法の適用は，ソースコードのみにとどまらない．本稿ではこの手法の画像データに対する適用もおこなった．多くの既存手法では，画像解析により対象画像から直接詳細な特徴を抽出している．しかし，このようにして得られた特徴量から全ての対象画像の特徴を統一された基準で表現するような低次元の射影を得るためには，高度な専門的な知識や膨大な試行錯誤が必要とされる．このため，人の意図に沿った類似性検出の実現は難しいと考えられる．人は目視で画像の類似度を判定するとき，無意識のうちに検索目的や嗜好などの何らかの指標に基づき画像のもつ全特徴 F_{all} から低次元の特徴ベクトル F_{sel} を抽出している．このとき，人の脳内には全ての画像が絶対的な尺度によりマッピングされているような特徴ベクトル空間が構築されているのではなく，既知の対象画像から抽出した特徴と新しく目にした対象画像の特徴を比較し，検索意図により適合する，もしくはより適合しないような小数の画像の特徴を代表値とし，これとの相対的な類似関係をもとに全体の類似度比較を行っていると考えられる．つまり，人はある時点での検索意図を関数として F_{all} から F_{sel} への写像を行っており，検索意図が変わるごとに関数は自動生成されていると考えることができる．本手法では，複数個の参照画像を用いて間接的に画像の特徴を表現することにより，人が自然に行っている類似性検出プロセスに倣った類似性検出を行う．参照画像は，先に述べた人の類似性検出プロセスにおいて，代表値の役割を果たす．参照画像を軸とする低次元特徴ベクトル空間上に F_{all} を写像することで，低次元の特徴ベクトルで人の検索意図に基づく特徴表現を行うことを目指す．この手法は参照画像の選択により特徴ベクトル空間を容易に再構築可能という点で柔軟な特徴抽出手法であるともいえる．

本稿では，実験により本手法がユーザの意図に基づく類似に近い画像検索が行えることが確認された．ただし参照画像が具体的に画像のどの特徴に対応しているのかについてはまだ明らかではない．

記述スタイルに基づく特徴抽出手法では，一次元データからこれまで表現されなかった特徴を抽出することを試みた．本稿においては，ソースコードから作成者の記述特徴を抽出し，記述スタイルモデルにより定量化する手法を提案した．これは，ソースコードそのものの特徴ではなく，ソースコードを記述する際に現れる，作成者固有の表記上の癖を記述スタイル特徴として抽出し，定量化するというものである．この手法では，ソースコードの構造や意味のような，ソースコードの大域的な特徴ではなく，作成者のコーディングス

タイルという，局所的な特徴を抽出するため，プログラミング授業における盗用発見に用いた際に，構造や意味に基づいた手法に比べて，偶然の類似による盗用の誤認識が発生しにくいというメリットがある．本手法は，人がソースコードを記述する際に見られるわずかな共通の傾向をモデルに学習させ，確率モデルにより表現するというユニークな立場をとっている．具体的には，作成者の記述スタイルを，隠れマルコフモデルをベースとした記述スタイルモデルで表現することにより，ソースコードにおける，ソースコードの内容には直接関与しない表面的な特徴を定量化する．本手法はチームでソフトウェア開発を行う際のスタイルチェッカーとしての適用や，ソースコードのアルゴリズムのみではなく記述スタイルを自動的に採点するという新しい試みなど様々な活用が期待されるが，その主たる適用として想定されるのはその作成者認識機能を活用したソースコード盗用問題の解決である．実験により，この手法は授業課題として提出されたソースコード長が短く内容が互いに類似したソースコード間において盗用を発見する手法として有効であることが確認された．人が授業課題ソースコード間の盗用を目視により発見しようとするとき着目する類似性を自動で高速に検出できることから，本手法は人の意図する類似性に基づいた特徴抽出手法であるといえる．

本研究では，人とコンピュータによる「協調型情報検索」を目指し，人の意図に基づく類似性検出を行うことができ，なおかつユーザ側で類似性尺度を容易に変更可能であるというような柔軟性を持つ i)参照を用いた特徴抽出手法，そしてこれまで実現されなかったテキストデータにおける見た目上の特徴を定量表現する ii)記述スタイルに基づく特徴抽出手法の 2 つを開発した．

この 2 つの手法をもとに，ソースコード間類似性検出手法，画像間類似性検出手法，そして記述スタイルによる授業課題ソースコード盗用発見手法という 3 つの実用的な手法を開発し，それぞれについて実験によりその有効性を確認した．

今後も本研究の目指す「協調型情報検索」の実現に向け，これまでに提案した手法のアルゴリズムの細部について更なる検討を行うと共に，新たな特徴抽出手法の開発や，それらを実装した実用的なシステムの開発を行っていく．

本論文の構成は，以下の通りである．

まず，第 2 章において本研究の背景である情報検索の現状と課題について述べる．次に，第 3 章では 2 つの特徴抽出手法，すなわち，参照を用いた一次元データの特徴抽出手法および記述スタイルに基づく一次元データの特徴抽出手法について説明する．第 4 章では参照を用いた手法の適用例として，ソースコードおよび画像データの類似性検出手法について説明する．続く第 5 章では記述スタイルに基づく手法の適用例として，授業課題ソースコードにおける盗用発見手法について説明する．最後に，第 6 章において，本研究全体の総括を行い，今後の課題について述べる．