様式 C-19

科学研究費補助金研究成果報告書

平成　23 年　5 月　16 日現在

機関番号：63801
研究種目：若手研究(B)
研究期間：2009 ～2010
課題番号：21700315
研究課題名（和文）　自然言語処理と集合知によるデータ解析手法の分類方法の開発
研究課題名（英文）　Development of a classification system for data analysis methods based on natural language processing and collective intelligence
研究代表者
小笠原　理（OGASAWARA OSAMU）
国立遺伝学研究所・生命情報・DDBJ 研究センター・助教
研究者番号：00435512

研究成果の概要（和文）：
昨今の測定技術の向上に伴い、生物学の分野などではデータドリブンな研究手法が注目されている。一方でコンピュータの高速化などに伴い、統計解析・データ解析手法は高度化し有用な解析手法が多数開発されている。これら 2 つの技術革新の融合はこれからの生物学研究に大きな影響を与えることが期待されるが、一方、大量データの測定・解析を行う実験研究者のような統計学の非専門家が先端的な解析手法にアクセスし正しく駆使することは容易ではない。

この状況を改善する目的で、私は遺伝解析手法のデータベース(R Graphical Manual)を 2006 年より公開してきた。関数の実行結果の画像を使って関数の機能を一望できるという特徴を持っており、2008 年の時点で月 10～50 万 page view, 月 8 千～1 万 unique IP ほどのアクセス数を持っており、世界中の研究者から利用され一定の評価を得ていた。しかしデータ更新に大きな計算量が必要であるにもかかわらず、サーバ環境やソフトウェアが十分整備されていなかった。

本研究において､このデータベースのサーバ環境､ソフトウェア環境を整えたことにより、2011 年 5 月の時点で月 20 万 page view, 月 5 万 unique IP となり、unique IP が顕著に増加した。月間 unique IP は DDBJ が 1 万 7 千、京都大学の KEGG が 20 万であるから、アクセス数については当初予想よりも大幅に増加し国内の著名なデータベースと比肩するサイトに成長した。

各種の統計学辞典や教科書および R Graphical Manual の関数マニュアルなどをもとに分類軸を作成した。この分類軸に R Graphical Manual 中の関数をマッピングする必要があるが、そのために R Graphical Manual の全文書に対して Named Entity Recognition (NER)を行い、統計学の専門用語を抽出し、それをもとにマッピングを行った。この目的で NER の精度を上げるために新しい方法を開発した。

研究成果の概要（英文）：
As data measurement technology has advanced, increasing attention has been paid to data-intensive approaches, especially in the field of biology. In addition, as the performance of digital computers has increased, so has the sophistication of statistical analysis and other data analysis methods. The fusion of data measurement and the data analysis technologies is expected to have profound impacts on future biological research. However, from a practical standpoint, it is difficult for experimental scientists who are devoted to making the measurements that generate massive amounts of data but are not specialists in statistics to make full use of cutting-edge statistical analysis methods.

To remedy the above-described problem, I have been publishing a database of statistical analysis procedures (the R Graphical Manual) since 2006. This database has the virtue that users can browse the functionality of procedures in the R statistical

system by making use of all the provided images generated by invoking all the examples in the R statistical system, as well as enabling full text search of all documents in the R statistical system. This database has been highly acclaimed by users world-wide and the visit statistics for the database were 100,000 to 500,000 page views/month and 8,000 to 10,000 unique IPs/month in 2008. However, sufficient resources, both in terms of hardware and software, had not been allocated to the database, despite the high computational demand necessary for data preparation for this database.

Thanks to the improved hardware and software environment of this project, the number of unique IPs per month has increased notably, to about 50,000 unique IPs/month (about 200,000 page views/month) in May 2011. Since the number of unique IPs/month of DDBJ (maintained by the National Institute of Genetics) is about 17,000 and that of KEGG (at Kyoto University) is about 200,000, the R Graphical Manual has grown in Japan into a database having comparable popularity to those famous databases.

In this project, I developed a classification system of statistical procedures taken from statistical dictionaries, textbooks, and manuals that are contained in the R Graphical Manual. In order to map the functions in the R Graphical Manual to the categories of this classification system, I developed a novel algorithm to improve the performance of named entity recognition (NER). This algorithm is applied to all the individual manual entries contained within the R Graphical Manual to extract technical statistical terms and I made a mapping from each procedure entry to the classification categories.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,400,000	420,000	1,820,000
2010 年度	600,000	180,000	780,000
年度			
年度			
年度			
総　計	2,000,000	600,000	2,600,000

研究分野：総合領域
科研費の分科・細目：情報学・統計科学
キーワード：データベース、自然言語処理、統計処理システム

１．研究開始当初の背景

昨今の測定技術の向上に伴い、生物学の分野などではデータドリブンな研究手法が注目されている。一方でコンピュータの高速化などに伴い、統計解析・データ解析手法は高度化し有用な解析手法が多数開発されている。これら 2 つの技術革新の融合はこれからの生物学研究に大きな影響を与えることが期待されるが、一方、大量データの測定・解析を行う実験研究者のような統計学の非専門家が先端的な解析手法にアクセスし正しく駆使することは容易ではない。

実験研究者が数多くの新しい解析手法を自分でプログラムに実装したのち膨大なデータを解析するということは現実的には難しい。統計学の専門家ではない研究者が豊富な統計手法を試してみる一つの現実的な方法は、専門家が作った統計手法のプログラムを使わせてもらうことであり、このようなプログラムが非専門家にもわかりやすい形で整理されていればデータ解析の際の現実的な手間は大幅に軽減される。これまで数学的概

念や方程式、統計手法のわかりやすい分類や解説を与えるサイトは MathWorld, PlanetMath, Encyclopaedia of Mathematics, EqWorld 等いくつか既に存在している。しかしそのようなサイトはプログラムそのものは提供しておらず、加えて編纂者が文書を書き起こす方式で作られているため、一般的で重要なことに注力して記載がなされている一方、最先端の手法や特定分野のみに適用される手法については収録されない傾向にあった。

この状況を改善する目的で、私は遺伝解析手法のデータベース(R Graphical Manual)を 2006 年より公開してきた。統計処理システム R は統計学の専門家の間でも人気が高く先端的なアルゴリズムが専門家により多数実装されており、それゆえ実験研究者の間でも人気が高いシステムである。私のデータベースは関数の実行結果の画像を使って関数の機能を一望できるという特徴を持っており、2008 年の時点で月 10～50 万 page view, 月 8 千～1 万 unique IP ほどのアクセス数を持っており、世界中の研究者から利用され一定の評価を得ていた。しかし、データベース作成にかなりの計算量を必要とするにもかかわらずデータ更新のためのソフトウェアシステムや検索システムが貧弱、サーバ環境も整っていないといった問題があった。

２．研究の目的

本研究の目的は、第一にすでに多数のアクセスがあり世界中の統計学者から認知されている R Graphical Manual のデータ更新、公開に関するハードウェア、ソフトウェア環境を整備することであった。第二に R で実装されているほぼすべての統計手法のパッケージおよび関数の分類を与えることにより、関連する統計手法を探しだすことを助けることである。

私が公開しているデータベースは、通常の文字列のキーワード検索のほかに、関数の実行結果の画像を使って関数の機能を一望できるという特徴を持っており、2008 年 10 月の時点で 1,300 個のパッケージ、33,000 以上の関数および 18,000 枚以上の統計解析グラフがデータベース化されていた。このように多数の関数が公開されているので、専門家ですら自分の分野に関係する関数にどのようなものがあるのかを把握することが困難であった。

この状況を改善するために、具体的には以下のことを目的に研究を行った。

(1) 私が公開している統計解析手法のデータベース(R Graphical Manual)の関数（現在約 6 万個）に対して分類を与える。これにより専門外のユーザでも必要な関数をより容易に探し出すことができるようにする。提案者自らが文献情報（教科書、辞書など）をもとに分類軸を作成する。

(2) 集合知の利用。集合知の形成を助けるためにさらなるアクセス数の向上を図る。そのうえでデータベースのユーザが分類を編集できるようにする。

(3) 関数の機能の説明文に対して類似文書検索を行う仕組みを実装する。

３．研究の方法

集合知が成立するためにはアクセス数が十分あることが大前提である。そのためにはデータ更新頻度を上げたり、検索の基本的な機能を整備するといった基本的な開発が必要であり、まず第一にこのような基本的開発を行った。

具体的には、本データベースはデータ更新の際、すべての実行可能な Example を実行し、中には計算量の多いシミュレーションなども含まれているので、データ更新の計算に１，２週間ほどかかる。データダウンロードからデータ更新、インデックスの作成の一連の作業を自動化するシステムを作成した。

次に、パッケージや関数の分類の作成についてであるが、人手による方法と、コンピュータによる計算を主にした手法の両面から研究開発を行った。

(1) 分類軸の作成 と分類軸に対するパッケージや関数のマッピング: 統計学の関数を分類するための分類軸（オントロジー）を教科書や辞書などの文献情報をもとに作成する。この分類軸に対して、R Graphical Manual のパッケージおよび関数をマッピングし、これにより統計解析関数の分類を行う。
(2) 集合知の利用 ： データベースのユーザが分類軸の編集やマッピングなどの追加・修正が行えるようなインタフェイスを作成する。集合知の成立のためにアクセス数の向上を図る。
(3) 自然言語処理技術による分類: 関数の機能の説明文を自然言語処理を用いて類似文書検索する仕組みを実装する。

４．研究成果

(1) 研究計画調書にある通り、集合知を集めるためにはアクセス数が一定以上あることが大前提である。以前は web 公開のためのサーバ環境すら十分でなかったが、本課題によりサーバ環境、更新プログラムの改善を行ったことでアクセス数は以前よりもさらに向上した。本研究を始めるときには、月間 unique IP が平均約 1 万程度であったが、現在は本体とミラーサイトの両方を合わせると平均 5 万程度に増加している。月間 unique IP は DDBJ が 1 万 7 千、京都大学の KEGG が 20 万であるから、アクセス数については当初予想よりも大幅に増加し国内の著名なデータベースと比肩するサイトに成長した。

(2) データベースが収録しているデータ量は、申請時 2008 年 10 月の時点で 1,300 個のパッケージ、33,000 以上の関数および 18,000 枚以上の統計解析グラフであったが、現在は 2,936 個のパッケージ, 58146 個の関数、36428 枚の統計解析グラフと、約倍の量に増えた。それに加えてソースコード検索や別の言語への拡張を行った。さらにユーザからの問い合わせに対応して検索の web API の整備を行った。

(3) 各種の統計学辞典や教科書および R Graphical Manual の関数マニュアルなどをもとに分類軸を作成した。この分類軸に R Graphical Manual 中の関数をマッピングする必要があるが、そのために R Graphical Manual の全文書に対して Named Entity Recognition(NER)を行い、統計学の専門用語を抽出し、それをもとにマッピングを行うこととした。この目的で NER の精度を上げるために新しい方法を開発した。本課題は、当初は自然言語処理のアルゴリズムの開発自体ではなく、既存の方法の応用として自然言語処理を駆使して関数の分類を行うことが主眼であった。しかし専門家に見てもらうレベルの分類とそこへのマッピングのために、その基礎となる Named Entity Recognition などの技術を開発するところから再検討した。この新しい計算方式については特許化も考えている。

(4) データの量が多いため、分類とパッケージや関数の間のマッピング作業は、計算機を使って行った。このマッピング結果が最適であるかどうかは、最終的には専門家の意見を反映させることが望ましい。サイトの利用者がマッピング結果を編集できるようなソフトウェアを開発した。

(5) 専門用語の共起関係を使った類似文書検索の web API を実装し、webAPI および web アプリケーションとして公開した。

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　0 件）

〔学会発表〕（計　0 件）

〔図書〕（計　0 件）

〔その他〕
ホームページ等

http://rgm2.lab.nig.ac.jp/RGM2/index.php
http://www.oga-lab.net/RGM2/index.php
(ミラーサイト)
http://rgm2.lab.nig.ac.jp/cgi-bin/awstats.pl
(アクセスログ解析結果)
http://www.oga-lab.net/cgi-bin/awstats.pl
(ミラーサイトのアクセスログ解析結果)

６．研究組織
(1)研究代表者
小笠原　理（OGASAWARA　OSAMU）
国立遺伝学研究所・生命情報・DDBJ 研究センター・助教
研究者番号：00435512