生命科学辞書とオントロジーの自動構築法の開発

●大久保 公策
国立遺伝学研究所　生命情報・DDBJ 研究センター　遺伝子発現解析研究室

＜研究の目的と進め方＞

ライフサイエンス分野の大量の文書 (文章) を内容に従って機械的に分類整理することを目的としている。生命科学は常に進歩を続ける分野であることを考慮しに全く新しい分野に対して辞書やシソーラスなど専門家が作った言語資源を必要としない新規技術の開発を目指す。

●分類整理には代表的である文書を用語の持つ意味ベクトルの和としてあらわし文書同士の関係を文書ベクトルの余弦であらわすベクトルスペースモデル（VSM）を用いる。

●独自の方法

VSM を用いる場合にすべての用語の意味を全く無関係であるとする用語数だけの次元を用いる場合、文書ベクトル同士のつくる直積は同じ用語の数に等しい共起用語の検出によるクラスタリングと同様である。　医科学文書の場合には 3 万種類以上の用語が数十ずつひとつの文書に存在するのが通常であるために、用語が共起する文書ペアーが少なすぎるためにほとんどの文書内容を表現できない問題がある。

これを克服するためにはあらかじめ用語同士の関係を大量の文書中で調べ、用語 x 教育用文書の行列で 0 のまますめを次元下げ近似の際の誤差として現れ数値を潜在的な用語と文書との関係として用いる LSI が知られている。(1990 Bell 研究所 Latent semantic indexing) LSI を行う場合の問題点では出来上がる用語関係が教育用の文書集合に影響されることにより、背景知識レベルの極めて自明な関係 (肝臓と肝細胞や肝臓が消化管と近縁にあること、癌はすべて共通の属性を持つことなど) が実文書 (たとえば PubMed) ではアンダーレプリゼントされていること。

この本来強く表現すべき基本知識を用語に与えるために我々が用いた方法は基礎医学系の生物学教科書を基礎知識として用いることである。具体的には教科書の Index セクションを利用して作ったページ x 用語行列を次元下げ近似することで LSI の効果を持つ用語ベクトルを用意し、この基礎知識用語空間に、LSI 法での質問文書と同様に内容不明の文書を投影する、すなわち文書中の教科書用語ベクトルの和として文書内容を表すというものである。

この際に教科書のページに対しては最適な内容の表現 (セクション見出しなど) が記載されているために、教科書の全ページとの余弦をつかえば質問文書の内容が用語の組み合わせという原始的なキーワード表現ではなく、医科学を構成するすべての代表的な話題との関係としてわかりやすい表現を与えることが可能となる。

＜ 2007 年度の研究の当初計画＞

1. 教科書の用語分布およびトピック配置を領域知識の形式的表現として利用し任意の文書内容を形式的に表現するシステムBOB

① 構築系および利用系の開発は現段階で一旦凍結し、論文で公開されている遺伝子発現データの解釈を疾患遺伝子発現に限定して数十例規模で行い、その評価(論文解釈との比較)検討を進める。

②上記を通じて、遺伝子機能の表現として連結したPubMedアブストラクト(Entrez gene)、数行の機能サマリー(Uniprot)、の双方について現在のBOBのパフォーマンスをある程度客観的に表現することを試みる。

2. 遺伝子発現整理およびヒトおよび動物ESTの生物種分類プログラムは昨年までに試験開発を終え開発利用を文部科学省統合データベースプロジェクトで実サービス前提の開発構築へ移行しました。

＜ 2007 年度の成果＞

これまでの開発内容は

・手作業での教科書目次および索引の入力、30Kレベルの用語集の整理、人手による用語集や索引データの校正。
・教科書追加に対応する上記辞書管理システムの構築。
・形容詞辞書および同義語辞書の構築。
・上記辞書の索引データへの適応による「意味ID」の「表現系リスト」「教科書：ページ座標データ」の管理DB構築
・自由に教科書を選択して用語ーページ行列（全データの部分行列）を選択し、SVDによるランク下げ近似表現を使ったLSIの作成プロセスの自動化。
・質問文書群を与えてパターン群を返すウエブサーバの開発。
・遺伝子列を与えてパターン群を返すウエブサーバの開発。
・用語パターンをコンコーダンスとして出力し、また指定ページの重みつき用語を返す教科書由来用語データの手作業校正補助システム開発。
・クラスター化遺伝子発現データを用いた解釈試験を通じた辞書データ中のエラーの検出と用語頻度重みのチューニング。
・全ユニジーンおよび全PubMedアブストラクトのプレインデクシングとページベクトルとの余弦のプレ計算による結果出力の高速化。
・GUI表示画像の分割保持による結果転送の高速化。

上記高速化法によって1,000程度の文書列を質問に与えたときに1分程度で2万あまりのページに対する余弦を表示することが可能になったため以下のように実データ試験および公開用のサービスの準備を行った。

・試験用（公開用）Webサーバーの構築。

1．教科書セクション名として日本語英語選択可能に。
2．ブックシェルフビューの改善
　パターンを比較する際のガイドとなるヘアカーソルの表示、サーマル表示機能を整備した。
3．複数教科書セクションの同時表示機能の追加

ブックシェルフビューで複数の教科書からセクション選択し、教科書ビューに結合して表示する機能を整備した。
4．オブジェクト情報表示機能の改善
オブジェクト（遺伝子、PubMed等）のテキストに対するタームマッピング情報の表示を高速化した。
5．IE7への対応
Microsoft Internet Explorer 7において動作するように改造した。
6．FireFoxへの対応　FireFoxにて動作するように改造した。
7．データの更新　RefSeq, PubMedの新規データの登録を行った。

以上によって完成した試験公開サーバーでは以下の動作を行う。
1．利用者がマイクロアレイやゲノム配列などによって得られた任意の遺伝子IDの序列を質問配列として与える。
2．質問IDをEntrexGene中での医学用語列の与えるベクトルとしてBOB中にマップする。
3．指示あるときはスペース内でクラスタリングを行う。
4．12冊の基礎医学教科書のすべてのページの余弦を計算。
5．12冊のすべてのページと質問ID列の余弦行列をヒートマップ行列として表示する。
6．表示から特徴的なパターンがあれば一冊の教科書を指定して拡大表示。
7．拡大表示中ではID列と教科書ページに対する項目見出しを読める。
8．関心部分のみ選択してレポートもしくは別の教科書へのマップに移行可能。

＜国内外での成果の位置づけ＞

遺伝子機能の形式的表現（オントロジー）を完全に自動的に作成することを意図した研究はみられない。

また教科書のページ序列および用語索引付けを利用して辞書形成から用語構造化を行う事例も生命系分野以外にも存在しない。

一方でGOなどの専門家の宣言によるツリー型の機能表現は非常に充実しており、本来生物種を超えた機能表現であるにもかかわらず高等脊椎動物に固有のシグナル伝達や形態形成などに関する機能についても充実を見せている。さらにマウスやヒトなど個別の生物における遺伝子のGOへの対応付けGeneOntologyAnnotationも常に更新をうける充実を続けており。事実上遺伝子機能の形式表現の世界標準となっている。

また任意の文書間の内容関係の表現は1992年のPubGene以来画期的な進展は見られず、遺伝子名称の検出に始まる遺伝子名索引付けデータからの構造化に終始している。

したがって遺伝子機能の形式表現と文書内容の表現については手法としての独自性に加えて以下のような相補的な特徴を備えている。
1．自動更新機能生命　科学の進歩を取り入れて更新するときに単純に教科書の大きな改訂を待てばよい、個別の遺伝子機能の発見への対応では遺伝子カード型のDBでの参考文献の付与を待てばよい。新しく発見された遺伝子に対する機能も該当文献を加えることですぐに対応可能である。

2．豊かな表現力　遺伝子機能やクラスターの意味表現にGOに用意されていない概念を選ぶことができる。すなわち20冊の教科書の200あまりのセクション見出しを答えのレパートリーとして保持している。

＜達成できなかったこと、予想外の困難、その理由＞
1．教科書由来の辞書の完全な訂正　辞書の改善は目視による方法しかないが、動議語が存在する場合でも表記が大きく異なれば目視で発見することが困難である。
2．用語ベクトルの重み付けの最適化　通常有名な遺伝子には数百におよぶ参考文献が付与されており、用語頻度分布は数百から1までの典型的なヘビーテイルを呈する。現在はTF-IDFによる古典的な重み付けしか行っていないがこれではほとんど数個の用語のみ利用しているのと同様の効果しかもたない。
3．結果の客観評価　そもそも解釈や分類は客観評価が不可能で主観評価にたよることになるがあらかじめ第３者が主観評価を行った正解セットは遺伝子発現のクラスター命名以外に利用できるものが見当たらない。

＜今後の課題＞

試験公開GUIを用いて　客観評価例を積み上げてその機能を第3者に受け入れられるものにする。

＜成果公表リスト＞

1）論文

1.0708101552
Hoshino H, Uchida T, Otsuki T, Kawamoto S, Okubo K, Takeichi M, Chisaka O.: Cornichon-like Protein Facilitates Secretion of HB-EGF and Regulates Proper Development of Cranial Nerves Molecular Biology of the Cell　Apr Vol.18 D1143-1152(2007).

2）データベース

1.0802251729
Okubo K,
BOB(試験公開サーバー)
http://222.151.240.6/project/bob/080208-/simple.cgi
ユーザ制限あるためにパスワード設定中