

汎用連想計算エンジンの開発と大規模文書分析への応用

Development of the generic association engine for processing large corpora

高野明彦, 丹羽芳樹, 西岡真吾, 岩山真, 久光徹, 今一修, 藤尾正和,
(NII) (日立製作所中央研究所)

徳永健伸, 奥村学 (東工大), 望月源 (北陸先端大), 野本忠司 (国文研)

1 はじめに

1.1 研究開発の背景

近年、百科事典や新聞を始め企業内文書まで、あらゆる文書情報の電子化が進み、それらの有効活用なしに充実した知的生活は考えられなくなっている。大規模文書コーパスを対象とする高速かつ高精度な情報検索技術が求められているが、現在実用となっているキーワード検索では、求めている文書が検索されなかったり、求めていない文書が大量に検索されるという問題がある。これらを解決するため、得られた文書集合の動的クラスタリング、文書集合に含まれる話題の自動要約、文書群を入力して類似文書を検索する文書連想検索などの有望な手法が提案されている。たとえば、我々が開発した文書連想検索と文書群要約機能を活用した検索インタフェースは、インターネット上での百科事典検索サービスで利用され始めている。しかし、これらの手法は文書群間・単語群間の類似性計算に基づくため、計算コストの高さから大規模文書コーパスでの本格的実用化には至っていない。

1.2 期待される効果、成果

大規模文書コーパスを対象とする文書連想検索や文書群要約は、統計的な類似性計算(連想計算)の計算コストが高いため、現在のところ精度と速度の両面で実用レベルには達していない。本研究開発は、これらの次世代情報検索技術の実現・評価に広く利用可能な高速の汎用連想計算ソフトウェアを作成し、この分野の研究者に共通の研究・評価基盤を提供することを目的とする。各種の類似性計算に共通に利用できる汎用で高速な連想計算エンジンが開発されれば、実用上意味のある大規模文書コーパスについて、提案されている各種手法の定量的な評価が可能となる。オープンソース形態で提供されるこのエンジンが、次世代情報検索技術実現のための共通基盤となり、新技術の実用化が大きく進展すると期待できる。本研究開発では以下の3項目の成果の達成を目指して来た。

- 汎用で高速な連想計算エンジン (ソフトウェア)

- 各種の連想計算を高精度かつ高速に実行する汎用連想計算エンジン。PC上で動作し、新聞1年分(約20万記事)に関する文書連想検索が2~3秒で可能となる。文書群同士・単語群同士・文書群-単語群間など各種の統計尺度を動的に切り替えて利用できる。Perl等のプログラムと連動させることにより、統計的計量に基づく文書や語彙の分析手法を簡便に実装でき、大規模文書集合についての実用性を定量的に評価できる。
- 文書集合の動的クラスタリング手法 (アルゴリズム+プロトタイプ)
 - 汎用連想計算エンジンを用いた高精度で高速な動的文書クラスタリング手法を開発する。それを高速文書クラスタリングのプロトタイプとして実装し、PC上で1,000文書のクラスタリングを2~3秒で可能とする。
- 計量的な語彙モデルに基づく語彙分析手法 (アルゴリズム)
 - 汎用連想計算エンジンの使用を前提に、語彙の計量モデルとその高速計算手法を開発する。それに基づく、大規模文書集合から特徴語(話題を担う語)の抽出手法や重要複合語の自動抽出手法について検討する。

2 研究開発の目標と内容

2.1 研究の目標

新聞や百科事典など文書数が数10万件を超える文書集合から、必要な情報をすばやく検索できる技術が強く求められている。現在実用となっているキーワード検索は、低い再現率(求めている情報が検索されない)と低い適合率(求めていない情報が大量に検索される)という問題を抱えている。これらを解決するため、我々はこれまで、検索時にキーワード集合ではなく文書それ自身を入力し、入力文書(群)と類似の文書を検索する文書連想検索 [2] を提案し、そのための文書クラスタリング手法 [1] を開発してきた。また、検索結果の内容的把握を助けるための文書や文書群の要約手法 [3, 4] を提案した。さらに、文書連想検索と自動要約機能の有機的連携を可能にする文書検索システム [4, 5, 6, 7] を自ら開発実装することを通じて、次世代情報検索技術の実用化の可能性を追求してきた。その結果、連想計算(文書群同士、単語群同士、文書群と単語群間の類似性関連性計算)こそが最も基本的で重要な計算機構であると確信するに至った。これらの研究的蓄積を踏まえ、本研究開発では、大規模な文書集合について各種の連想計算を高精度かつ高速に処理できる汎用連想計算エンジンを開発する。また、作成した汎用連想計算エンジンを研究上のツールとして使い、動的クラスタリング・要約手法、計量的語彙モデルに基づく語彙分析手法について研究を推進する。さらに、それらの研究成果を利用して、汎用連想計算エンジンの改良を図る。

- 汎用連想計算エンジンの開発
 - 文書クラスタリング、文書連想検索、特徴語抽出の計算で必要だった文書群同

士・単語群同士の類似性計算を拡張して、各種の統計的計量に基づく連想計算モデルを作成する。文書群を単語のマルチセットで、単語群を文書のマルチセットでモデル化する場合のように、数学的に双対(Dual)の関係にある2つの統計的計量が同時に必要とされることが多い。この一組の双対な計量についての連想計算が単一のインデックスを用いて同等に高速処理できるよう(双対な)データ構造を設計する。文書クラスタリングでは、各文書は文書数が1件の文書群(クラスタ)と見なされるので、文書と文書群を同等に扱えるデータ構造設計が必須となる。このような性質を備えた基本データ構造を高度に圧縮して保持する高速な汎用連想計算エンジンを作成する。

● 動的クラスタリング・要約手法の開発

上記の汎用連想計算エンジンを用いて、精度の犠牲なしに文書クラスタリングや文書連想検索を高速に処理する手法を検討する。高速文書クラスタリングの実現により、検索結果の自動分類や、検索結果の要約を検索者に提示するなど、高度な検索インターフェイスが可能になる。本研究項目では、オンラインで高速に動作する高速文書クラスタリングのプロトタイプを汎用連想計算エンジンを用いて実装する。また、検索結果の動的な分類、および要約生成の可能性についても検討する。要約生成にあたっては、語の話題性に関する語彙分析手法の成果の利用を検討する。検索結果のみならず文書群の適切な要約は、入力文書数が極端に多い文書連想検索にとって必須の技術と考えられる。

● 計量的語彙モデル、語彙分析手法の開発

情報検索において単語の果たす役割は想像以上に大きい。人間は語という短い文字列から、自らの知識や経験に基づく多くのものを意識的無意識的に想起できる。我々はこの観点から、検索結果の文書群を特徴づける単語集合の自動抽出方法 [4, 5, 6, 8] を提案してきた。また、「語はその出現環境により特徴づけられる」という考え方に基づき、語の(相対)頻度や特定の構文パターンにより語の重要度を計量してきた [9]。本研究ではこれを発展させる。

[語彙の計量モデル] 文書集合中の単語の意味を「文脈を共有する語の分布＝共起語分布ベクトル」などの統計的計量として扱うことを検討する。上記汎用連想計算エンジンを用いて語彙の計量モデルの計算手法を開発する。

[語彙分析手法] この計量モデルを使って、文書集合中での語の「話題性」の統計的な評価法を考案する。さらに、複数の語が複合することにより話題性が格段に強まる語が「重要複合語」であると考え、それらの自動抽出法を提案する。計算コストの高い共起関係解析に汎用連想エンジンを用いることにより、新聞1年分や百科事典などの大規模文書集合についても、PCレベルの計算資源で上記の語彙分析が可能であることを実証する。また実際に話題性の高い語を集中的に提示することにより、検索支援機能が向上することを実証する。

2.2 ベースとして用いる理論、技術

我々はこれまで、階層的ベイズクラスタリング (HBC) [1] とそれに基づく文書連想検索 [2] を提案してきた。また、文書集合に含まれる話題の要約を、特徴的な単語群とそれらの共起関係グラフとして示す手法 (特徴語グラフ) [4] を提案し、さらにこれらを組み合わせた検索インタフェース DualNAVI [7] を開発した。汎用連想計算エンジンの基本計算モデルとしては、これらの手法の核となっている各種の統計的類似性計算を一般化したものを用いる。

2.3 研究開発のスケジュール

本研究は平成11年度から開始し、平成13年度における終了を予定している。これまでの研究開発成果および本年度 (平成13年度) 当初の研究開発計画は以下の通りであった。

平成11年度

- 汎用連想計算エンジン (第1版) の開発
- 連想計算に基づく次世代情報検索技術の調査研究
- 計量的語彙モデルの調査研究

平成12年度

- 汎用連想計算エンジン (第2版) の開発
- 高速文書クラスタリング・プロトタイプシステムの開発
- 計量的語彙分析手法の調査研究

平成13年度

- 汎用連想計算エンジン (第3版) の開発
- 高速単語重要度計算プロトタイプシステムの開発
- 連想計算に基づく情報アクセス手法の調査研究

3 平成13年度の活動状況

3.1 汎用連想計算エンジンGETA (第3版) の開発

これまで平成11~12年度には、各種の統計的計量に基づく連想計算の高速実行が可能で、使用する統計的計量を動的に変更できる汎用連想計算エンジンGETAの開発を行ってきた。またその性能を損なわずに連想計算エンジンの基本機能を簡便に利用できるPerlインタフェースや高速クラスタリング・ライブラリ等の実験環境の提供を行ってきた。

本年度、平成13年度は、1,000万件規模の文書コーパスへの適用を目的とし、高性能計算機として一般化しつつあるPCクラスタ上で動作する分散処理方式を設計し、任意規模のPCクラスタで利用可能なGETA (第3版) を開発した。

主な開発内容は以下の通りである。

(1) PCクラスタ上での分散型連想計算の基本設計

- (2) 各クラスタでの分散連想モジュール
- (3) 複数の分散連想計算を管理するモジュール
- (4 a) T C P / I P 対応の通信モジュール
- (4 b) M P I 対応通信モジュール
- (5) 分散型連想計算用データの作成・分配モジュール

また、G E T A を評価するためのシステムの設計開発も併せて行った。

3.2 高速単語重要度計算プロトタイプシステムの開発

前年度は文書中での単語の話題性を評価する手法に関して調査検討を行い、統計的計量 (representativeness) を求める手法を提案した。本年度はそこで得られた知見に基づいて以下の2項目の開発・研究を行った。

基本ツールのライブラリ形式化

計量をG E T A を用いて計算する際の基本ツールをライブラリの形で提供し、利用方法に関する文書を作成した。主な開発項目は以下の通りである。

- (1) 単語の話題性原指標計算ライブラリの基本設計
- (2) 文書集合に対する話題性原指標計算モジュールの設計・開発
- (3) 単語ブール結合に対する話題性原指標計算モジュールの設計・開発
- (4) 利用者用マニュアルの作成

重要複合語の自動判別化

語が複合することにより話題性が高まる「重要複合語」の自動判別手法を検討した。

- (1) 重要複合語の自動判別方法の調査・研究
- (2) 複合語の孤立性に基づく重要性判定方式の開発

3.3 連想計算に基づく情報アクセス手法の調査研究

G E T A の利用により初めて高速処理可能となった「文書クラスタリング」「語彙連鎖解析」「文書群を特徴づける単語群の抽出」等の基本的な文書分析手法をベースに本年度は以下の2項目を実施した。

「文書クラスタリング」を用いた新情報アクセス手法の調査研究

- (1) 情報アクセス支援における文書クラスタリングの利用に関する調査研究
 - 適合性フィードバックへの応用手法
 - 増進的フィードバック手法との比較検討
 - 検索要求をバイアスとして反映させる手法
- (2) 検索結果のクラスタリングを用いた直観的インタラクションモデルに関する調査研究
 - クラスタリング結果の木構造で表示する方法
 - クラスタリング結果を色彩的に表示する方法
 - 検索性テストコレクションを用いた評価

「文書群を特徴づける単語群の抽出」を用いた新情報アクセス手法の調査研究

- (1) 情報アクセス支援における「特徴単語群の抽出」の利用に関する調査研究
 - 検索結果から特徴語を動的に抽出する手法
 - 特徴語間の関連性を抽出する手段
- (2) 「文書群を特徴づける単語群の抽出」を用いた直観的インタラクションモデルに関する調査研究
 - 特徴語を視覚的に配置・表示する方法
 - 検索結果と特徴語の並列表示を利用したインタフェース
 - 特徴語の表示を利用した適合性フィードバック

4 外部発表および成果物

4.1 外部発表論文

1. Hisamitsu, T., Niwa, Y., Nishioka, S., Sakurai, H., Imaichi, O., Iwayama, M. and Takano, A. Extracting Terms by a Combination of Term Frequency and a Measure of Term Representativeness. *Terminology*, 2001.
2. Hisamitsu, T. and Niwa, Y. Topic-Word Selection Based on Combinatorial Probability. *Proceedings of NLPRS2001*, pp.289-296, 2001.
3. Iwayama, M. Relevance Feedback with a Small Number of Relevance Judgments: Incremental Relevance Feedback vs. Document Clustering. *Proceedings of SIGIR2000*, pp.10-16, ACM Press, 2000.

4. Iwayama, M. and Niwa, Y. and Nishioka, S. and Takano, A. and Hisamitsu, T. and Imaichi, O. and Sakurai, H. and Fujio, M. The Effect of Document Clustering in Interactive Relevance Feedback *Proceedings of NTCIR-2 Workshop*, pp.196–203, 2001
5. Niwa, Y., Iwayama, M., Hisamitsu, T., Nishioka, S., Takano, A., Sakurai, H., and Imaichi, O. DualNAVI - dual view interface bridges dual query types. *RIAO2000 Innovative Applications*, 2000.
6. Niwa, Y., Hisamitsu, T., and Imaichi, O. Similarity based Approach for Filtering Important Word Bigrams. *Proceedings of SNLP2000*, pp.195–205, 2000.
7. 岩山真. 適合性フィードバックの効率化について, 情報処理学会情報学基礎研究会, 2001-FI-57, pp.1–8, 2001
8. 岩山真, 丹羽芳樹, 西岡真吾, 高野明彦, 久光徹, 今一修, 櫻井博文, 藤尾正和. 対話的文書検索における文書クラスタリングの役割, 情報処理学会情報学基礎研究会, 2001-FI-62, pp.129–136, 2001
9. 久光徹, 丹羽芳樹. 共起単語中の特異的頻出単語数を用いた用語の representativeness 計測尺度, 情報処理学会第63回全国大会論文誌, 3H-04, 2001
10. 高野明彦, 丹羽芳樹, 西岡真吾, 久光徹, 岩山真, 今一修. DualNAVI による連想的テキストアクセス電子情報通信学会ソサイエティ大会、チュートリアル企画「発見科学とデータマイニングの最前線」, 予稿集, 2001
11. 西岡真吾, 今一修. 汎用連想計算エンジン GETA に基づく連想検索システム LACE. 日本ソフトウェア科学会 WISS2000 デモセッション, 2000.

4.2 成果物

- (1) 「汎用連想計算エンジンの開発と大規模文書分析への応用」研究成果報告書
(ソフトウェアのソースコード1式を含む)
 - ・汎用連想計算エンジン GETA (第3版) の開発
 - ・高速単語重要度計算プロトタイプシステムの開発
 - ・連想計算に基づく情報アクセス手法の調査研究
- (2) 汎用連想計算エンジン GETA (第3版) 基本設計書
- (3) 汎用連想計算エンジン (第3版) 導入・操作マニュアル

5 おわりに

5.1 まとめ

平成11～12年度は、各種の統計的計量に基づく連想計算の高速実行が可能で、使用する統計的計量を動的に変更できる汎用連想計算エンジン GETA の開発

を行ってきた。またGETAの高速性能を損なわずに連想計算エンジンの基本機能を簡便に利用できるPerlインタフェースや高速クラスタリング・ライブラリ等の実験環境を提供した。この段階で当初計画で目標として来た連想計算エンジンはほぼ完成した。

本平成13年度は、(当初目標には含まれていなかったが)電子化文書量の指数関数的増大に鑑み、1,000万件規模の文書コーパスへの適用を目指して、高性能計算機として一般化しつつあるPCクラスタ上で動作する分散処理方式を設計し、任意規模のPCクラスタで利用可能なGETA(第3版)を開発した。分散の度合(用いるCPUの数)に応じた計算性能の測定も行い、文書量が増大すればする程、分散版GETAが必要かつ有効であることが明らかになった。

本年度はその他、高速単語重要度計算プロトタイプシステムの開発を行い、前年度に提案を行った統計的計量(representativeness)を求める手法に基づき計量をGETAを用いて計算する際の基本ツールをライブラリの形で提供した。また重要複合語の自動判別化に関しても、中心となる類事例探索プログラムを作成した。

また連想計算に基づく情報アクセス手法の調査研究を行い、GETAの利用により初めて高速処理可能となった「文書クラスタリング」「語彙連鎖解析」「文書群を特徴づける単語群の抽出」等の基本的な文書分析手法をベースにして、「文書クラスタリング」を用いた新情報アクセス手法の調査研究と「文書群を特徴づける単語群の抽出」を用いた新情報アクセス手法の調査研究を行い、調査報告書に示したような成果を上げた。

5.2 今後の課題

本年度をもって、本テーマは終了したが、大規模情報を扱う上での連想計算は不可欠なツールであるとの実感をさらに深めている。今後は本テーマで開発したツールを、広く研究その他の目的に役立ててもらえるように最大限の努力をしたいと考えている。当面の施策として、以下を早急に実施する予定である。

- ユーザシンポジウムの開催
- Web上における、GETA Homepageの立ち上げ、およびツール群の公開。

GETA ホームページについては、国立情報学研究所からの発信を予定している。

参考文献

- [1] M. Iwayama and T. Tokunaga. Hierarchical Bayesian Clustering for Automatic Text Classification. *IJCAI'95*, pp.1322-1327, 1995.
- [2] 岩山, 徳永. 確率的クラスタリングを用いた文書連想検索. *自然言語処理*, Vol.5, No.1, pp.101-117, 1998.
- [3] T. Nomoto, Y. Matsumoto. Data Reliability and Its Effects on Automatic Abstracting. *Fifth Workshop on Very Large Corpora*, ACL-SIGDAT, 1997.

- [4] Y. Niwa, S. Nishioka, M. Iwayama, A. Takano, and Y. Nitta. Topic Graph Generation for Query Navigation. *NLPRS'97*, pp.95–100, 1997.
- [5] Y. Niwa, M. Iwayama, and A. Takano. Interactive Support of Query Refinement by Dynamic Word Co-occurrence. *ICCPOL'97*, pp.383–386, 1997.
- [6] 丹羽. 動的な共起解析を用いた対話的文書検索支援. **情報処理学会研究報告**, Vol.96, No.87 (96-NL-115), 1996.
- [7] 西岡, 丹羽, 岩山, 高野. 文献検索支援インタフェース DualNAVI. **日本ソフトウェア科学会 WISS'97**, 1997.
- [8] Y. Niwa and Y. Nitta. Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries. *COLING'94*, pp.304–309, 1994.
- [9] T. Hisamitsu and Y. Niwa. Extraction of Useful Terms from Parenthetical Expressions by Using Simple Rules and Statistical Measures. *COMPUTERM'98*, pp.36–42, 1998.