一般化文書頻度の計数ライブラリおよびシステムの開発

― 大規模文書集合の統計量分析・検索システム ―

1. 背景

1980 年代後半から，自然言語処理の研究分野では言語データ(コーパス)に基づくデータ主導型アプローチという研究の流れが起こり，統計的手法が多く用いられている．他方で，情報検索の研究分野においてもまた統計的手法が主に用いられている．

これらの応用で利用される統計量は，主に単語の頻度が基本となる．例えば，文字列が文書集合(コーパス)中に出現する総数である文字列(索引語)頻度や，文字列が出現する文書数である文書頻度がよく利用されている．従って，大量の言語データからこれら文字列の頻度に基づく統計量を効率的に計算し，またそれを高速に参照することは，基盤となる重要な技術であるといえる．

他方で，筆者らは文字列の頻度に基づく統計量の一種であり，文書頻度をより一般化した一般化文書頻度を提案している．一般化文書頻度 $df_k(P)$ は文字列 P が k 回以上含まれている文書の数であり，文字列 P の意味に関連する性質を持つ有益な統計量である．

Church は，$df_2(P)/df_1(P)$ により推定可能な反復度(adaptation)という特徴量を導入し，反復度が単語の頻度よりも語彙に強く依存することを示した[2]．武田らは， 表 1 に示すように，反復度が語の境界で大きく変化することに着目しキーワードの自動抽出に応用した．加えて，これを情報検索に利用し，その有効性を示した．

表 1 一般化文書頻度と反復度の例

文字列	df_1	df_2	df_3	反復度
メ	52424	22324	11117	0.426
メデ	4632	2200	1221	0.475
メディ	4580	2178	1211	0.476
メディア	4434	2131	1195	0.481
メディアを	560	88	15	0.157
メディアを用	83	12	0	0.145
メディアを用い	83	12	0	0.145
メディアを用いた	64	6	0	0.094

2. 目的

本プロジェクトの目的は，任意文字列を分析の対象として，一般化文書頻度を 10GB～100GB という大規模な文書集合(コーパス)について獲得するためのライブラリとシステムを開発することである．

3. 開発の内容

3.1. システムの機能と特徴

本プロジェクトにて開発するシステム(以降，単にシステムと呼ぶ．)の概要を図 1 に示す．システムは，任意の文字列 P が与えられたとき，コーパス D における文字列 P の一般化文書頻度 $df_k(P)$ $(1 \leq k \leq K)$ を計算し，それを応答として返す．ここで，K は予めシステムに与えられている求める重複度の上限である．また，副次的に，文字列 P のコーパス D における出現頻度である文字列頻度 $tf(P)$ や，文字列 P のコーパス D における出現位置なども得ることができる．

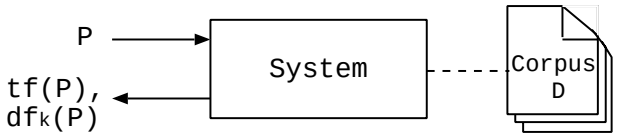

図 1 システムの概要

システムには以下の 3 つの大きな特徴がある．

1. 辞書中の単語やバイグラムではなく，任意の文字列 P を対象とする．
2. 文書頻度 $df(P)$ だけではなく，それをより一般化した一般化文書頻度 $df_k(P)$ を得ることができる．
3. 10〜100GB 程度の大規模コーパスを対象とする．

この中で，システムの最大の特徴は 10〜100GB とい大規模なコーパスを一般化文書頻度計数の対象とすることで，従来の手法の限界が数 100MB であったのに対し，10〜100 倍程度の改善を実現した．

3.2. システムの構成

システムでは，任意の文字列 P の一般化文書頻度 $df_k(P)$ を高速に計算するために，予めコーパスの全部分文字列について $df_k(P)$ を求め，その結果をデータベースとして格納しておく．一般化文書頻度 $df_k(P)$ は，このデータベースを検索することによって得る．すなわち，システムは図 2 に示すように，コーパス D からデータベース DB を構築する DB_BUILDER モジュールと，文字列 P の問い合わせに対して，データ

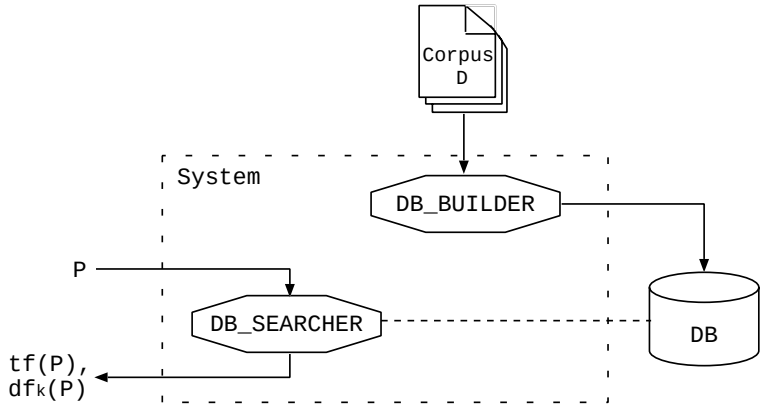

図 2 システムの構成

ベース *DB* を検索することで $df_k(P)$ を求め，応答として返す DB_SEARCHER モジュールの 2 つから構成される．加えて図 3 に示すように，複数の PC に構築したデータベースを蓄積し，それらを並列に検索し，結果を統合することによって大規模コーパスに対応している．

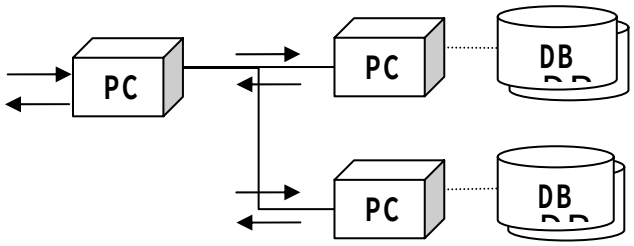

図 3 分散検索

3.3. システムの性能

主記憶容量 1GB の PC を用いて、コーパスサイズを変化させながらデータベース構築を行った結果を図 4 に示す．従来の手法(old)が主記憶容量の 1/10 が限界であるのに対し，開発したシステムでは(new)，ほぼ主記憶容量と同じサイズのコーパスについてデータベースを構築できる．

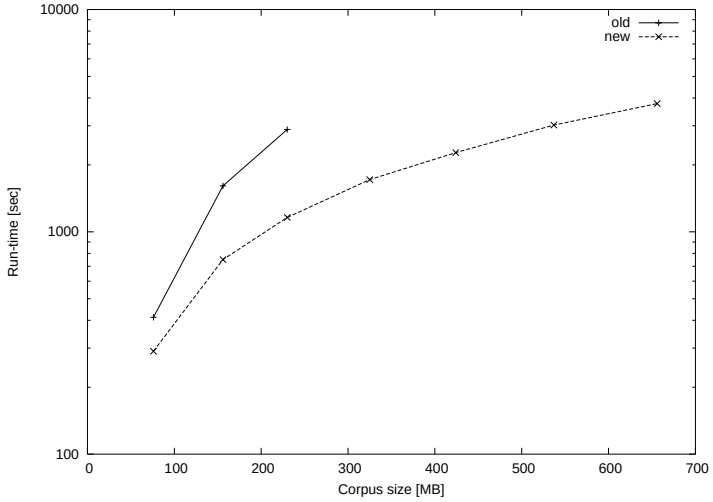

図 4 データベース構築の実行時間

4. 従来の技術との相違

任意文字列に対して一般化文書頻度を高速に取得するアルゴリズムは既に文献[1]で提案されており，本プロジェクトでも文献[1]の手法をベースにしている．しかし，このアルゴリズムは一般化文書頻度の分析・検索のためにコーパスサイズの約 *10* 倍もの主記憶容量を必要する．このため，一般的な PC では数 *100*MB 程度のコーパスの分析が限界であり，スケーラビリティの点で問題があった．本プロジェクトでは，この従来手法の問題点を解決し，これらの大規模コーパスに十分対応可能なスケーラビリティを有したシステムを開発した．

5. 期待される効果

従来は数 100MB が限界であった全部分文字列に対する一般化文書頻度の計数が 10GB～100GB という大規模コーパスに適用可能となったことにより,一般化文書頻度の応用範囲が大きく広がることが期待できる.例えば,Web や特許を対象としたコーパスは非常に巨大でこれまでは一般化文書頻度の計数が困難であったが,これらの大規模なコーパスに対して統計量の分析が行えるようになり,反復度を用いたキーワードの自動抽出や,それを利用した自然言語をクエリとした情報検索を行うことができる.

また,ライブラリ・システムとして整備されることは,一般化文書頻度自身の研究にとっても有益である.一般化文書頻度や反復度の性質はいくつかの論文で詳しく調べられているが,文書頻度や文字列頻度の性質の分析に比べてまだ十分とはいえない.システムを利用することにより,一般化文書頻度の性質の分析をスムーズ実施できる.

6. 普及・活用の見通し

開発したシステムの活用としては,文献[3]に示されている一般化文書頻度に基づく辞書を全く用いることのないキーワード自動抽出や,それを利用した自然言語をクエリとして使用することができる情報検索のシステムが挙げられる.

7. 開発者名

寺尾健一郎

（豊橋技術大学大学院 情報工学専攻 teraken@ss.ics.tut.ac.jp）

参考文献

[1] 梅村恭司,真田亜希子. 文字列を k 回以上含む文書数の計数アルゴリズム. 自然言語処理,Vol.9,No.5,pp.180-186,2002.

[2] Kenneth W.Church. Empirical estimates of adaptation:The chance of two noriegas is closer to $p/2$ than p^2. In Coling-2000, pp.180-186,2000.

[3] Yoshiyuki TAKEDA, Kyoji UMEMURA, and Eiko YAMAMOTO. Determining Indexing Strings with Statistical Analysis. IEICE Trans. on Information and Systems, Vol.E86-D, No.9, pp.1781-1787, 2003.