D-37

大規模検索システムにおける概念辞書自動更新

Auto Updating of a Concept Dictionary in Large Scale Search System

相川 勇之† 伊藤 山彦† 高山 泰博† 鈴木 克志†
Takeyuki Aikawa Takahiro Ito Yasuhiro Takayama Katsushi Suzuki

1．はじめに

インターネットの普及とともに企業内文書の電子化が進み，大量の情報から必要な情報を峻別するための検索技術が重要になっている．我々は，表記が異なっていても類似の内容をもつ文書を検索可能な概念検索技術を利用した情報検索システムを開発している．

代表的な概念検索方式として，各文書における単語の出現傾向をベクトル情報として表現し，ベクトルの類似が文書の意味の類似を表すものとして類似文書を検索するベクトル空間モデルがある．ベクトル空間モデルに基づく概念検索方式は多数発表されているが，我々は大量文書から単語の類似性を学習して得られる概念辞書を介して類似文書を検索する方式[1][2]を採用している．

本方式には，文書索引サイズの増大が登録文書数に対して線形でおさえられるという利点がある．一方，概念辞書に登録されていない単語からは検索ベクトルを生成できないという課題がある．そのため従来の検索システムでは，運用開始後に出現する新出単語への対処としては，定期的な概念辞書の保守以外にはなく，運用コストが大きくなる．

この課題を解決するため，登録履歴から新出単語を検出し，新出単語の概念ベクトルを計算して概念辞書を自動更新する機能をもつ，保守が容易な大規模検索システムを提案する．本稿では，提案システムの構成及び概念辞書の学習方法について説明し，概念辞書自動更新アルゴリズムについて述べる．

2．概念検索システム

この節では，提案システムの構成と概要について述べる．図1は，提案システムのブロック構成図である．

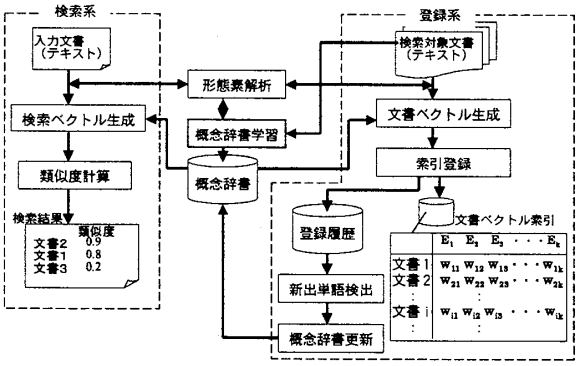


図1　概念検索システムの構成

提案システムは，検索系と登録系からなる．検索系では，入力された検索文を形態素解析して単語に分割し，概念辞書を参照して得られる各単語の概念ベクトルから検索ベクトルを合成する．概念辞書はあらかじめ登録対象文書から学習する．学習方法については第3節で説明する．

登録系では，登録対象文書中のテキストを同じく形態素解析により単語に分割し，各単語の概念ベクトルより文書ベクトルを合成する．合成の際には，各単語に対する tf・idf 重みを適用する．概念辞書および文書ベクトル索引は固定次元であるため，文書数増大に対する索引サイズの増加を線形で抑えることができるという特徴をもつ．

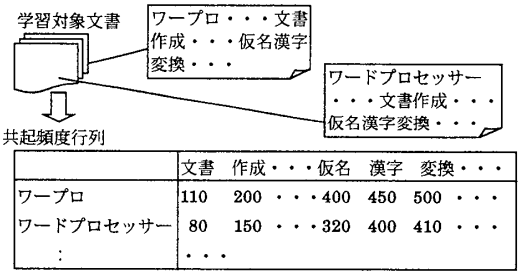
さらに索引登録と同時に，新出単語検出のために，一定期間の登録履歴を記録する．登録履歴から新出単語を検出し，検出された新出単語の概念ベクトルを計算して概念辞書に追加登録する．登録時に新出単語への対応を自動的に行なうことで，概念辞書の定期的な保守作業を不要とする．

3．概念辞書の学習

この節では，概念辞書の学習方式について説明する．

3.1 Word Space モデル

われわれは，概念辞書学習のモデルとして word space[1][2]を採用している．word space とは，単語の共起頻度行列（図2）を特異値分解により圧縮して得られる概念空間である．



	文書	作成・・・仮名	漢字	変換・・・
ワープロ	110	200 ・・・400	450	500 ・・・
ワードプロセッサー	80	150 ・・・320	400	410 ・・・
：	・・・			

図2　共起頻度行列

この方式では，単語の意味的な性質をその単語と共起する単語の統計として定義する．例えば，「ワープロ」と「ワードプロセッサー」は「文書作成」「仮名漢字変換」などの単語との共起傾向が類似するので，類似の意味をもつとする．

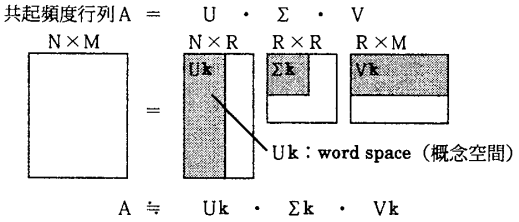


図3　特異値分解の適用

一般に共起頻度行列は疎な性質をもっているが，単語数が数万から数十万に及ぶと，非常に大規模な記憶空間を必要とするため実用的ではない．そこで共起頻度行列がもつ疎な性質を利用し，図3に示した特異値分解を適用することにより元の共起頻度表の性質を近似した概念空間に変換して記憶容量を圧縮する．図3において，N行M列の共起

† 三菱電機株式会社 情報技術総合研究所
Information Technology R&D Center, Mitsubishi Electric Corp.

頻度行列Aは特異値分解により一意にU，Σ，Vの3項の積として分解できる．UはN行R列の行列であるが，このうち固有値の大きい順に k 次元をとったUk・Σk・Vk により，元の共起頻度行列Aを近似できる．提案システムでは文献[2]と同様に，Ukを概念空間として使用する．

3.2 概念辞書学習アルゴリズム

概念辞書の学習処理は以下の3段階で行なう．

(1) 共起頻度収集

まず，学習対象文書を形態素解析し，単語に分割する．さらに同一段落中に共起する単語の頻度を集計し，共起頻度行列を作成する．大量の文書を学習対象とした場合，最終的に生成される共起頻度表は非常に巨大な表となるため，一次記憶上では処理しきれない．二次記憶に随時出力し，最後にこれらを集計する．その際に共起頻度の大きい情報のみ残し，高速化および省メモリ化を行なう．共起頻度の閾値は次項の切り出し処理に十分な情報量をもつ範囲となるよう設定する，

(2) 部分行列の抽出

特異値分解には大量のメモリ空間を必要とする．現実的な処理コストで特異値分解を実行するため，頻度上位の単語を処理対象とする．(1)で得られた共起頻度行列からN語×M語の部分行列を切り出して使用する．N及びMについては，学習対象文書の規模に応じて決定する．

(3) 特異値分解

上記の(2)で得た共起頻度部分行列に対して特異値分解を実行する．特異値分解の結果得られる k 次元の3つ組（Uk，Σk，Vk）のうち，左特異ベクトルUk を概念辞書として使用する．各単語に対する概念ベクトルは長さが1となるよう正規化する．上記3つ組のうち，Σk およびVk については，次節で説明する概念辞書の自動更新処理において，概念ベクトル逆演算用のデータとして使用する．次元数 k は検索精度と速度性能及びメモリ性能とのトレードオフで決まるが，数十万文書規模の場合，300次元～1000次元程度あれば十分な性能が得られる．

4. 概念辞書の自動更新

この節では，概念辞書の自動更新アルゴリズム，および特許明細書を対象とした予備実験について説明する．

4.1 登録履歴の記録

検索システム運用開始後の文書登録処理において，登録文書の履歴を記録する．登録履歴は一定期間のみ保持することとし，古い記録から破棄していく．登録履歴を保持する期間については，記憶コストと概念ベクトル計算精度とのトレードオフとなるが，最低限の統計情報を得るため数万文書規模の登録履歴を保持するものとする．

4.2 新出単語の検出

登録履歴から新出単語を検出する．検出には，概念辞書の学習における共起頻度表切り出しと同じ頻度基準を用いる．具体的には，概念辞書学習時の文書数を D_{All}，登録履歴中に蓄積された文書数 D_{hist} としたとき，式(A)で定義する検出基準頻度 F_{nw} 以上の単語を新出単語として抽出する．

$$F_{nw} = N \times D_{hist} / D_{All} \quad \text{式(A)}$$

4.3 新出単語の概念ベクトル計算

上記(2)で抽出した新出単語のそれぞれについて，登録履歴中での共起情報を収集する．収集結果として，（新出単語n語）×（共起単語M語）の共起頻度行列 A_{nw} が得られる．ここで共起単語M語とは，概念辞書の学習で部分行列の切り出しに用いた語数である．

こうして得られた A_{nw} に対して，特異値分解の結果得られたΣkおよびVkを用いた近似手法である folding-in[3]により，概念ベクトルを計算する．総登録文書数が数万件以下と少なければ近似手法を用いるかわりに概念辞書を再学習することも可能だが，登録文書数が数十万件以上の大規模な検索システムでは，概念辞書の再計算は処理コストが大きすぎて現実的ではない．

4.4 予備実験

新出単語数 n が学習時の語数Nに対して十分に少ない場合に，新出単語に対する概念ベクトルが有効な近似となることを期待できる．文献[4]では LSI(Latent Semantic Indexing)における folding-in の適用実験がなされており，近似を適用する新規登録行列の行数が，特異値分解を適用する元行列の行数の 30%を越えると近似の歪により検索精度が悪化することが報告されている．ここで新規登録行列は本研究の A_{nw} に相当し，元行列はAに相当する．

実データにおける新出単語数を調べるため，1996 年度から 2001 年度までの公開特許公報6年分を対象として以下の予備実験を行なった．まず，各年度の明細書を形態素解析し，各年度において出現頻度 100 以上かつ出現文書数 20 以上の自立語を抽出した．これらを辞書に追加登録していくことを仮定し，累計単語数と新出単語数の推移を算出した（図4）．

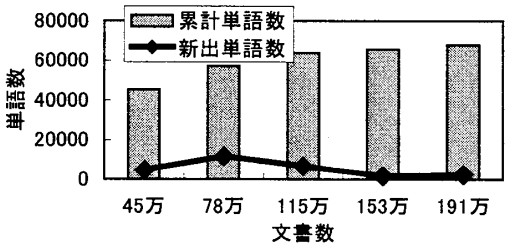


図4 累計単語数と新出単語数

上記抽出条件における 2001 年度の新出単語数は約 2300 語であり，それまでの累計単語数約 66000 語となる．この条件においては新出単語数は少なく，folding-in の近似による検索精度低下は小さくなることが期待できる．

5. まとめ

概念辞書を自動更新する機能をもつ，保守が容易な大規模検索システムを提案した．今後は，提案手法の有効性を確認するための実験及び評価を実施する予定である．

参考文献

[1] H. Schütze, J. O. Pedersen：A coocurrence-based thesaurus and two applications to information retrieval, Information Processing & Management, Vol 33, No3, pp.307-318, 1997.

[2] 高山他：単語の連想関係に基づく情報検索システム InfoMAP，情報処理学会研究会 情報学基礎 53-1，1999.

[3] M.W. Berry, et al.：Computational Methods for Intelligent Information Access, Proceedings of Supercomputing'95, 1995.

[4] H.G. Zha, et al.：On Updating Problems in Latent Semantic Indexing, Technical Report No. CSE-98-002, Pennsylvania State University, 1998.