# 文書関連性に基づく検索モデルの提案

3V-07

金沢 輝一† 高須 淳宏‡ 安達 淳‡

† 東京大学大学院工学系研究科 ‡ 学術情報センター研究開発部

## 1 はじめに

情報検索において, 問い合わせ表現 (query) をシステム側で補うことで検索性能を向上させることができる. 代表的な手法として query expansion [1] が挙げられるが, 元々情報量の小さい query を高い精度で拡張することは難しく, 補われた語彙によって不要な文書が検索される率が高まることがある.

本論文では文書関連性を用いて文書ベクトルを拡張することで検索性能を向上させる手法を提案する. 提案手法では検索テーブル作成時に文書関連性に基づいて文書集合を作り, この集合を単位として補う要素を決定することで精度の向上を図る. 評価実験として, 学術論文に予め付与されているキーワードを情報源として文書間の関連性を抽出, この関連性を元にベクトルを補って, tf-idf との性能比較を行う.

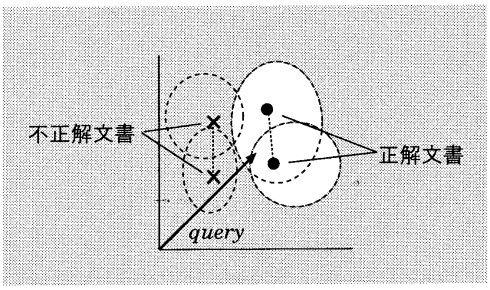
## 2 意味的曖昧性の問題

文書検索処理においては, query として自然文あるいは語の列挙の形で入力を行う方法が一般的だが, query は検索者の意図を代表する表現の一つに過ぎず, 表現とそれによって示される概念は一対一対応とは限らないために, query から検索者の意図を正確に汲み取ることが難しい場合もある. これは意味的曖昧性の問題と呼ばれており, 情報検索においては実用的な検索精度を得るために曖昧性への対策が必要となる.

代表的な手法として query expansion が挙げられるが, query という元々情報量の小さい要素に対する処理であるために, 再現率が向上する一方で適合率が犠牲になるという問題が生じる.

## 3 文書関連性を考慮した検索手法の提案

本論文では query expansion とは対照的に, 検索対象の文書間に存在する関連性に基づき文書集合を作り, これを解析することで文書側でベクトルを拡張する手法を提案する (図 1).




■ 図 1 文書ベクトルの拡張 (提案手法)

query expansion と比較した場合, query という情報量の小さい要素を拡張するよりもベクトル拡張の精度が高まるので, 適合率の低下を抑制しながら再現率を向上させることができると予想される.

## 4 評価実験

提案手法によるベクトル拡張の効果を確認するために実験を行った.

### 4.1 検索対象データ

- NACSIS において整備が進められている IR 用日本語テストコレクション [3] のテスト版を利用.
- 検索対象は学会発表 DB から 33 万 9501 件分となっており, 各発表に関して
  - ・発表題目
  - ・要約文
  - ・発表者が付与した自由なキーワード (複数)

  が記録されている. 今回の実験では題目と要約文を検索対象とし (以下「文書」と呼ぶ), 付与キーワードが完全一致するものを関連性がある文書とみなす.
- 題目と要約文は表記の修正後に形態素解析プログラム ChaSen[2] による解析を行い, 体言・用言のみを検索対象語として抽出した.
- 今回の実験では文書ベクトルは検索対象語 $t_j$ の出現頻度を $j$ 番目の要素とした. 以下, ここで求めたベクトルを**基本ベクトル**と呼び, 文書 $d_n$ の基本ベクトルの $j$ 番目の要素を $d_j^n$ と表す.

---

Retrieval Model Based on Relevance of Documents.
Teruhito KANAZAWA†, Atsuhiro TAKASU‡, Jun ADACHI‡
† Graduate School of Engineering, Univ. of Tokyo
‡ R & D Department, NACSIS

- テストコレクションの正解は

  ランク **A** … 正解. 検索要求にレレバント.

  ランク **B** … 部分的な正解. 検索要求に部分的にレレバント.

  という基準が設定されており, 暫定的な判定によって 32 件が利用可能となっている.

### 4.2 文書ベクトルの拡張

1. 今回の実験ではキーワードは完全一致したものを同一とみなし, 同じキーワードが付与されていることをもって文書が関連性を持っていると判断する. データベース中にはキーワードが 381781 種存在し, 複数の文書に付与されていたものは 133201 種であった.
2. キーワード $k_l$が付与されている文書群 $\mathcal{K}_l$の代表ベクトルを作る. 代表ベクトルの要素は $\mathcal{K}_l$に含まれる文書のベクトル要素の算術平均とした. すなわち, $\mathcal{K}_l$の代表ベクトル$\boldsymbol{K}_l$の $j$ 番目の要素は

$$k_j^l \equiv \frac{1}{|\mathcal{K}_l|} \sum_{\mathrm{d}_n \in \mathcal{K}_l} d_j^n \tag{1}$$

3. 次のように文書ベクトルを拡張する.

$$d_j'^n \equiv \max(d_j^n, k_j^0, k_j^1, ..., k_j^m) \tag{2}$$

ただし, $k_j^0, \cdots, k_j^m$は文書 $\mathrm{d}_n$が属す文書群 $\mathcal{K}_0, \cdots, \mathcal{K}_m$の代表ベクトルの第 $j$ 要素である. 以下, $d_j'^n$によって構成される文書ベクトルを修正ベクトルと呼ぶ.

### 4.3 検索実験の概要と評価方法

実験は次のようにして行なう.

- 基本ベクトルによる tf·idf と提案手法の性能を比較する.
- 自然文で与えた query から形態素解析によって文書ベクトルと同様に体言・用言を抽出する. query の各語の重みは 1 に固定した.
- query ベクトルと文書ベクトルの内積によって文書の順位付けを行う.

一方, 再現率・適合率の計算を,

$$(\text{再現率})R_i = i/(\text{正解文書数}) \tag{3}$$

$$(\text{適合率})P_i = i/(\text{上位から } i \text{ 番目の正解文書の順位}) \tag{4}$$

によって行う.

また, 次の式による積分適合率を定義し, 各手法・各 query ごとに計算する. 32 件の query による積分適合率の平均を手法の性能とする.

$$P_{\mathrm{int}} \equiv \frac{1}{N} \sum R_i P_i \tag{5}$$

### 4.4 実験結果と考察

表 1 に実験結果を示す.

■ 表 1 積分適合率による性能比較

| | 正解集合 A | A+B |
|---|---|---|
| tf·idf | .1892 | .1768 |
| 提案手法 | .2043 | .1891 |
| ポイント差 | +.0151 | +.0123 |
| 性能向上率 | +7.98% | +7.95% |

提案手法はランク A の正解集合でも A+B でも 8%程度適合率が向上した. これはランク A の文書, ランク B の文書, 不正解文書という順に得点が付けられているという望ましい状況を示している.

問題点を挙げると, 関連語ではあるが文書の内容を表してはいない語が補われ, 性能向上を阻害している場合がある. 当面は

- 付与キーワードが完全一致の文書だけを関連性があるとしている点.
- キーワードの代表ベクトルの要素を算術平均としている点（式 1）.

この 2 点に注目しての改良を考えている.

## 5 おわりに

本論文では自然言語の意味曖昧性が情報検索において問題となっていることを取り上げ, 文書関連性を用いて文書ベクトルを拡張することで検索性能を向上させる手法を提案した. また, 実験により手法の有効性を示した.

## 参考文献


[1] Chris Buckley, Amit Singhal, Mandar Mitra, "Using Query Zoning and Correlation Within SMART : TREC 5," 1996.

[2] 日本語形態素解析器 茶筌 (ChaSen), http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html

[3] "NTCIR : NACSIS Test Collection Project [Poster]," Kando,N., Koyama,T., Oyama,K., Kageura,K., Yoshioka,M., Nozue,T., Matsumura,A. and Kuriyama,K., *the 20th Annual Collquium of BCS-IRSG*, March 25-27, 1997, Autrans, France.