3U-02

代表ドキュメントによる検索精度向上効果の測定

武田 善行　　山本 英子　　梅村恭司
豊橋技術科学大 情報工学系

1. はじめに

複数のキーワードの組み合わせによる完全一致型の検索システムについて考える．このような検索システムを使って検索を行う場合ユーザは次のような手順で検索を行う．

・検索したい事例に関連する語を幾つか抽出する
・その中でも検索に有効と思われる候補を選ぶ

このような検索の形態では，このユーザからの入力の時点で検索に有効な情報の損失が起こっている．本来 query として与えた単語を含む事例の検索を行うべきところで，単語の区切りとインデクシングを意識しなければならない．ここで失われた情報は復元不可能なものである．

また，SIGIR-97 の論文の中で，バージニア大学の Fox らは，インターネットでは一般的な検索文は 1 つか 2 つの単語であると言っている[1]．このように，キーワードを組み合わせたタイプの検索システムに対する query は，一般的に極めて短く，情報量が少ない．このような条件下では，過去に検索に有効であるとされたテクニックの多くは，効果的な適用が不可能である．

このようなシステムにおいて検索要求からは十分な情報を得ることができないということがわかった．そこで，より高い検索性能を得るためには，何らかの形でそれを補う必要がでてくる．本研究では，コーパスからの統計的な情報を元に，query expansion とは異なるフレームワークとして，その補完を行い，特に確からしい事例に対する検索性能の向上をねらう．

2. 代表ドキュメントの利用

本研究での検索のアプローチはコーパスに現れる類似した事例の集中を利用しようというものである．本研究ではこの，集中した事例のうち，最も代表的なものを代表ドキュメント[2]と呼ぶ．代表的なものとする尺度は数多く存在し，また応用によって適切なアルゴリズムも異なるものであるが，本研究では，最も仲間を多く持つ事例を代表とする．キーワードだけで選ばれたドキュメントから代表を求めれば，検索性能が向上すると考えた．

3. 意味に踏み込んだ類似度

ここで，どのようなアルゴリズムを使っても，代表を選び出すときには「意味」が似ているドキュメントの数を判定する必要がある．さらにいえば，「意味」が似ていることを判定する類似尺度が不適切であった場合，どのようなアルゴリズムを使用しても，正しい代表を選ぶことはできない．

このような背景から，本研究では，「意味」が似ていることを判定することを念頭においた類似尺度を用い，キーワードが含まれるか否か以上の情報の獲得を目標とした．

4. DP 法

本研究では前述した要求を満たす類似尺度として DP 法を採用した．DP 法の定義は次のようになる．

[定義]

α, β, ξ, η を文字列とする．α_{km} を k 番目の文字から $k+m-1$ 番目の文字までの α の部分文字列，α_{n*} を n 番目の文字から最後の文字までの α の部分文字列とする．β_{km} を k 番目の文字から $k+m-1$ 番目の文字までの β の部分文字列，β_{n*} を n 番目の文字から最後の文字までの β の部分文字列とする．また，$Score$ は文字列から実数値を求める関数とする．

$$SIM_{DP} \max_{i,j}(Comp(\alpha_{1i}, \beta_{1j}) + SIM_{DP}(\alpha_{i+1*}, \beta_{j+1*}))$$

但し，$Comp(\xi, \eta)$ は次のように定義される．

Improving the Precision of retrieval results by obtaining typical documents
Yoshiyuki Takeda, Eiko Yamamoto and Umemura Kyoji
Toyohashi University of Technology Dept. of Information and Computer Sciences

・ $if\ \xi = \eta\ then\ Score(\xi)$

・ $if\ \xi \neq \eta\ then\ 0.0$

但し $Score(\xi) = -\log_2(df(\xi)/N)$

5. 代表を求める

少数のキーワードによって検索された結果は，多くの同じスコアのドキュメントを含むと考えられる．その中で，検査結果の代表となりうるものを優先することにした．y をキーワードを含むクラスタの要素，C を似ているか否かの閾値とすると，代表を求める式は次のようになる．

$$\max_x \arg n\{SIM(x,y) > C\}$$

6. query expansion との比較

query expansion[3]は，検査結果の上位のドキュメントを質問の中に取り込み，限られた情報を補うとする手法である．本文で提案する方法も，これと同じように限られた情報を補う手法であるが，ドキュメントの空間で情報を補おうとしているところが異なる．類似したものに文献[4]のようなものがある．さらに，ドキュメントの空間での類似度について，情報検索で用いられる類似度ではなく，意味に踏む類似度を使っているところが特徴である．

7. DP 法の優位性

DP 法が既存の類似尺度(BN:ngram 法)に比べ,意味に踏み込んでいることを示すために，表現の揺れに対し，類似度がどれだけ追従するかという傾向を観察した．

表 1 は，順位 1 の文字列に関して，自分を含む 10 個の文字列(順位 1 の文字列の表記を揺らしたもの)と，それぞれの類似度を BD 法，BN 法により算出したものである．比較のために，類似度は同一文字列を与えた場合との相対値としてある.順位は DP 法の結果に従って示した．

これによると，表記が変わることにより，BN 法の類似度が急激に落ち込むのに対し，DP 法の類似度が緩やかに下降しているのがわかる．よって，BN 法に比べ DP 法が意味に踏み込んだ検索を行えることがわかる．

表 1 DP 法は BN 法より意味を反映する
Table 1 DP is more effective than BN for meaning.

順位	DP	BN	文字列
1	1.000	1.000	機械翻訳システムの出力を人間が編集することが必要である.
2	0.843	0.625	機械翻訳のシステム出力を人間が編集することが必要である.
3	0.708	0.561	機械翻訳結果を人間が編集することが必要である.
4	0.699	0.397	機械による自動翻訳システムの出力は,人間が編集することを必要とする.
5	0.696	0.536	機械による翻訳のシステムのあとに人間が編集することが必要である.
6	0.695	0.658	機械翻訳システムの出力を人間の手を加えることが必要である.
7	0.689	0.877	機械翻訳システムの出力を編集することが必要である.
8	0.617	0.531	コンピュータによる翻訳出力を人間が読んで,編集することが必要である.
9	0.407	0.245	機械翻訳システムの問題点は,そのままでは使えず,後編集が必要なことである.
10	0.373	0.109	機械翻訳したあとに人間の編集作業が必要である.

8. システムの構築状況

現在，システムはプログラミングが終了しており，代表ドキュメントを求めることができている．その類似度においては，章 7 で示されているように，通常使われる類似度よりも意味に踏み込んだ類似度になっていることが確認されている．情報検索の問題を使った評価は現在作業中である．

References

[1] Gerald Kowalski: Information Retrieval Systems Theory and Implementation, KLUWER ACADEMIC PUBLISHERS(1997)

[2] 山本 英子, 武田 善行, 梅村 恭司: 代表的なドキュメントを求めるための類似尺度, DEWS, 6A-4(1999)

[3] Chris Buckley, Gerard Salton, James Allan, and Amit Sigha : Automatic query expansion using smart, Trec 3. In The Third Text Retrieval Conference(TREC-3), pp. 69-80(1995).

[4] 金沢 輝一, 高須 淳宏, 安達 淳: 関連性の重ね合わせモデルによる文書検索, DEWS, 5B-5(1999).