ISS-P-030

KL 情報量を用いた情報検索

樋山　友理香†　　三浦　孝夫†
† 法政大学理工学部創生科学科

1. はじめに

近年，インターネットの普及に伴い日々膨大な文書データが生成され，ユーザが意図している文書抽出が重要となっている．現在，情報検索において類似度を求める手法は，主に余弦波類似度を用いる．しかし余弦類似度では，質問と検索対象である文書のうち片方のみ出現している単語を考慮しないという問題がある．この解決手法としてカルバックライブラー情報量(KL 情報量)があるが，内容が類似していない文書で使用される単語数が少ない場合，その文書も抽出される．本稿では，KL 情報量を用いた使用単語数に依存しない検索手法を提案する．

2. 提案手法

KL 情報量を用いて，使用される単語数を考慮した類似度の指標を提案する．aを質問文書の見出し語数，bを検索対象の文書における見出し語数，cをaとbの双方に生じる見出し語数とするとき，式を下記のように示す．

$$\frac{(a+b-c)}{(c+1)} * \frac{D_{KL}}{(c+1)} \qquad (1)$$

短い文書では使用される見出し語数が少なくなるため，D_{KL}の値が小さい場合でも第1項の値が大きくなり，全体として値が大きくなる．見出し語数が質問文書と検索対象の文書で同じならば，第1項の係数は1となり KL 情報量と同じになる．各項の分母は，$c = 0$の場合を考慮し，補正する．

4. 実験

4.1 実験手順

本研究では 2 つ実験を行う．まず KL 情報量において，見出し語数の少ない文書が高い類似度になることを回避可能か実験する．そのため，同じデータを用いて一般の KL 情報量と提案手法の比較を行う．質問ベクトルと文書ベクトルの KL 情報量を求め類似度を算出する．その結果得られた類似度の上位20件にあたる各語彙数とランキングが提案手法後で何位に移るかを評価する．2つ目の実験として，質問と検索対象である文書のうち，片方のみ出現している単語を考慮した情報検索の実験を行う．ここでは検索手法として余弦類似度と提案手法の比較をする．第1の実験と同様に，各手法の類似度からランキングを求める．上位30位まで抽出した文書内容と質問文書の内容から，真に類似している文書であるか評価する．

4.2 実験データ

本稿では CD-毎日新聞2017年版に採録されている1月1日から2週間分のデータを使用する．1記事を1文書としたとき，文書数は2334となる．また，質問文書は2334文書を除いたものからランダムで1つ選択する．次に検索に用いる単語を形態素解析により抽出し，Zipf の第二法則を用いることで低頻度の単語を削除する．

その結果得られた単語は 2442 語，質問文書に含まれる延べ語数は133語である．これらを用いて，tfを適用した文書ベクトルと質問ベクトルを生成する．

4. 3. 結果

KL 情報量と提案手法での類似度ランキングの一部の変化を以下に示す．

表1　KL 情報量と提案手法のランキングの一部

総単語数	KL情報量での順位	提案手法での順位
13	11	1834
9	12	2185
7	13	2288
21	14	1027
42	15	838
6	16	2416
26	17	1864
25	18	822

本来内容が類似せず，総単語数の少ない文書が KL 情報量では上位に存在するが，提案手法では類似している文書ではないとして下位に位置している．これはD_{KL}の値は小さいが，式(1)で示した第1項で$a \gg b$という条件により全体の値が大きくなっていると考えられる．これより，KL 情報量では短い文書で見出し語数が少ないものが上位になるのを回避できている．

次に余弦類似度と提案手法の類似度上位30位までの文書内容を表2に示す．ランダムで抽出した質問文書は将棋の結果の記事であったため，検索対象の文書が将棋の内容を含んであるものを正解とする．その検索における文書内容の正解数・誤答数と，お互いの手法で検出している文書で抽出できていない文書数を表3に示す．

表3　文書内容の内訳

	余弦類似度	提案手法
正解数	24	26
誤答数	6	4
片方が出せてない数	1	2

提案手法は余弦類似度より2記事多く正解数を抽出でき，余弦類似度にない正解文書も取得できている．これより提案手法の方がより情報検索において有効である．

5 結論

KL 情報を用いて，見出し語数を考慮した類似度指標の提案を行った．双方に生じる単語を考慮することにより短い文書のランキングを上位1%から最小でも35%下げられた．また，余弦類似度よりも高いランキングで2記事多く類似文書を抽出し，提案手法の有効性を示した．

(参考文献)高村大也: 言語処理のための機械学習入門、コロナ社、2010