F-022

Earth Mover's Distance を用いた記事関連度計算方式における高速化手法
Speedup Method in Measuring the Degree of Association between Articles with EMD

八尾 学人† 吉村 枝里子‡ 土屋 誠司‡ 渡部 広一‡
Manato Yao Eriko Yoshimura Seiji Tsuchiya Hirokazu Watabe

1. はじめに

近年電子化情報の急激な増加により，ユーザが入手可能な情報は膨大なものとなり，利用者の要求に合った情報を的確に探し出す必要性が高くなっている．そのためには検索要求と検索記事との間の類似性や関連性を定量化することが求められる．その際，従来の情報検索では記事における単語の出現頻度など，統計情報を利用して検索要求と検索記事間の類似性を判断し，記事を選別[1]している．このような手法は検索要求と記事内の各単語の表記が一致しない場合は関連性がないとの仮定に基づいている．だが実際の記事において単語間は互いに意味的な関連性を持っている．そこで概念ベース[2]と Earth Mover's Distance（EMD）を用いて記事と検索要求の類似度を計算する手法[3]を用いる．しかし問題点として既存の記事関連度計算方式と比べると精度は良いが計算時間が非常に長いことが挙げられる．そのため，本研究では EMD による記事関連度計算を高速化するアルゴリズムを提案する．

2. 関連技術

2.1 概念ベース

概念ベースとは複数の電子国語辞書から機械的に構築された大規模な知識ベースである．ある単語を概念として定義し，その意味特徴を表す語である属性と，属性の重要度を数値で表した重みの対の集合によって構成されている．概念は約 9 万概念登録されている．例を表 1 に示す．

表 1　概念ベースの例

概念	属性，重み
医者	(医者, 0.72) (患者, 0.58) (病気, 0.20) ...
患者	(患者, 0.74) (病院, 0.46) (病状, 0.10) ...
治す	(治療, 0.41) (薬, 0.10) (病気, 0.07) ...

2.2 関連度計算方式

関連度計算方式[4]とはある 2 つの概念間の関連の強さを定量的に表現する手法である．関連度は 0.0 から 1.0 までの実数値で表現され，概念間の関連が強いほど大きな値を示す．概念間の関連度を計算する際には各概念の属性間にどれくらい一致する属性があるかを示した一致度を用いる．

一致度とは，概念 A,B で表記一致する属性の小さい方の重みの総和である．小さい方の重みを用いるのは，両概念の属性に共通して存在する重み分は有効だと考えるためである．概念 A,B の一致度は以下の式で表現される．

$$DoM(A,B) = \sum_{a_i=b_j} \min(u_i, v_j) \quad (1)$$

概念 A,B の一次属性をa_i, b_j，重みをu_i, v_jとし，各概念の持つ属性の個数を L 個，M 個とすると，概念 A,B は，以下の式で表現される．

$$A = \{(a_1, u_1), \ldots, (a_L, u_L)\} \quad (2)$$

$$B = \{(b_1, v_1), \ldots, (b_M, v_M)\} \quad (3)$$

概念 A,B の関連度は以下の式で表現される．

$$DoA(A,B) = \sum_i DoM(A,B) \times \frac{u_i+v_j}{2} \times \frac{min(u_i,v_j)}{max(u_i,v_j)} \quad (4)$$

3. EMD を用いた記事関連度計算方式

記事関連度計算方式とは関連度計算が語と語の関連の強さを表現するのに対して，記事関連度計算方式は記事と記事の関連の強さを定量的に表現することができる．定量化された値を記事関連度と呼ぶ．

EMD とは分布間の距離を表すもので，最適な輸送コストを用いて定義される．EMD を記事検索に適用する場合には，需要地と供給地，需要量と供給量，各需要地と供給地間の距離を定義する必要がある．需要地には検索要求の索引語を，供給地には検索記事の索引語を割り当てる．ここでの索引語とは記事の中の数詞を除く名詞，動詞，形容詞である．

需要量と供給量はそれぞれの索引語の重み，需要地と供給地間の距離は索引語間の関連性と見立てることができるため，索引語の重みを輸送すると考えて，重み配分は各索引語間の一致度を考慮して総輸送量が最小になるようにしたものである．例として「河川敷でお花見」と「公園の桜で宴会」の関連度を算出する．図 1 に記事検索の適用例と表 2 にその時の一致度を示す．また図 1 中の括弧内の数値は各索引語の重み，点線の矢印が使用する一致度であり，矢印の下の数値は使用する一致度である．

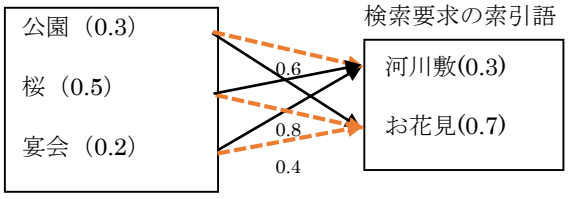

図 1　EMD を用いた記事検索の適用例

表 2　記事検索適用例の一致度

	公園	桜	宴会
河川敷	0.6	0.1	0.2
お花見	0.3	0.8	0.4

一致度と索引語の重みから輸送量(重み×(1-一致度))が少なくなるように求める．最小にした総輸送量を検索要求の重みの和で正規化して EMD を算出し，輸送量が大きいほど関連は低いので，1-EMD を記事関連度とする．以下の図 2 に計算の具体例を記載する．

†同志社大学大学院理工学研究科
Graduate School of Science and Engineering, Doshisha University
‡同志社大学理工学部
Faculty of Science and Engineering, Doshisha University

桜⇒ お花見：0.5×(1-0.8)=0.1
公園⇒河川敷：0.3×(1-0.6)=0.12
宴会⇒お花見：0.2×(1-0.4)=0.12
総輸送量：0.1+0.12+0.12=0.34
EMD：(0.1+0.12+0.12)÷1.0=0.34
記事関連度：1-0.34=0.66

図 2　EMD を用いた記事関連度の計算

4. 提案手法

既存手法で EMD を求める際の輸送問題の解の算出手法は 2 段階シンプレックス法[5]である．しかしシンプレックス法は計算数が膨大で記事関連度計算に非常に時間がかかる．そこで輸送問題を解く上で輸送問題の特徴である制約条件が全て n 元 1 次方程式であるという特徴を活かした手法を導入する．初期解は Houttaker 法[6]と北西隅法[6]を用いる．最適化は MODI 法[6]と飛び石法[6]を用いる．輸送問題に特化した解法であるほうが計算数は少なくなるため，EMD を用いた記事関連度計算方式の高速化が実現できると考えられる．初期解 2 つと最適化 2 つの組み合わせの 4 種類を提案手法として評価と計算時間を求めた．北西隅法と飛び石法を提案手法 1，Houttaker 法と飛び石法を提案手法 2，北西隅法と MODI 法を提案手法 3，Houttaker 法と MODI 法を提案手法 4 とする．

5. 評価方法

EMD を用いた記事関連度計算方式の評価として NTCIR3-Web[7]からの引用により検索要求 36 件と検索記事 1000 件を使用する．各検索要求に対して検索記事は適合の具合が H 判定（高適合），A 判定（適合），B 判定（部分的適合），C 判定（不適合）の 4 段階であらかじめ設定されている．本研究は H 判定，A 判定を適合記事としてそれぞれの検索記事に対して平均精度を求めそれらの 36 件の平均をとったものを平均精度の平均を精度とする．平均精度(AP)と平均精度の平均(MAP)は以下の式で求められる．

$$AP = \frac{1}{s}\sum_{i=1}^{n}\frac{z_i}{i}\left(1+\sum_{k=1}^{i-1}z_k\right) \quad (5)$$

$$MAP = \frac{1}{T}\sum_{h=1}^{T}AP_h \quad (6)$$

i：順位
z_i：順位 i の記事が適合するなら 1 適合しないなら 0
S：適合記事の総数
n：出力記事数
T：検索記事
AP_h：検索記事 h に対する平均精度

6. 評価結果

平均精度の平均(MAP)と計算時間を表 3 に示す．

表 3　評価結果

	MAP	計算時間(s)
提案手法 1	0.375	350236
提案手法 2	0.375	42048
提案手法 3	0.375	551
提案手法 4	0.375	340
既存手法	0.386	84555

提案手法 4 は既存手法に比べて MAP が 0.011 低下したが，計算時間では約 84210 秒早くなり既存手法と比べて大幅に短縮することができた

7. 考察

提案手法を用いることで計算時間を短縮することに成功した．しかし MAP に関しては誤差の範囲ではあるが低下した．Houttaker 法は北西隅法より初期解を導出するのに時間がかかるが最適解に近い初期解を導出ができる．MODI 法は飛び石法にくらべて 1 回の計算には時間がかかるが，1 回の計算でより最適解に近づく．つまり結果から少しずつ最適解に近づけるよりも，1 回の計算に時間がかかっても一度に最適解に近づけるほうが計算時間は短くなるといえる．

精度向上の課題点として以下の 2 つがあげられる．1 つ目は固有名詞への対応である．固有名詞は概念ベースに登録されていないため属性を取得することができず，一致度の計算が行えない．例として検索課題「宮部みゆきの執筆した小説に対する書評・レビューが読みたい」の AP は 0.0106 だった．概念ベース以外の言語資源を用い，固有名詞の判別及び意味理解を行う必要があると考える．2 つ目は索引語の取得法である．索引語の取得には形態素解析機である茶筌[8]を用いているが適切に分けられない場合がある．例として，「スレッド」という言葉を茶筌にかけると「ス」と「レッド」という様に分かれてしまう．形態素解析の結果にカタカナの語は 1 つの形態素とするなどのルールで手を加えることが必要であると考えられる．

8. おわりに

本稿では索引語の関連性を概念ベースにより定量化し，EMD の考え方を取り入れて記事の関連性の求める手法に輸送問題の特徴に着目して高速化手法を提案した．結果として計算時間は大幅に短縮することができ，また検証を行うことによりこれからの改善方法を見出すことができた．今後の研究課題は精度の向上である．概念ベースに登録されていない語への対応や形態素解析手法の見直しをすることにより精度の向上が期待できるのではないかと考える．

謝辞

本研究の一部は，科学研究費補助金（若手研究（B）24700215）の補助を受けて行った．

参考文献

[1] 倉田篤史，渡部広一，河岡司，“概念ベースと関連度計算方式を用いた記事関連度計算方式”，情報処理学会研究報告，2006-NL-171，pp.19-24，2006.
[2] 奥村紀之，土屋誠司，渡部広一，河岡司，“概念間の関連度計算のための大規模概念ベースの構築”，自然言語処理，Vol.14, No.5，pp.41-64，2007.
[3] 藤江悠吾，渡部広一，河岡司，“概念ベースと Earth Mover’s Distance を用いた文書検索”，信学技報，Vol.108，No.456，pp.111-116，2009.
[4] 渡部広一，奥村紀之，河岡司，“概念の意味属性と共起情報を用いた関連度計算方式”，自然言語処理，Vol.13，No.1，pp.53-74，2006.
[5] 向譲治，“2 段階シンプレックス法” “http://zeus.mech.kyushuu.ac.jp/~tsuji/java_edu/TwoPhase.html” (参照 2014-7-11)
[6] 前田活郎，“オペレーションズリサーチ”，朝倉書店，1973.
[7] NTICRProject“NTCIR|HOME”http://research.nii.ac.jp/ntcir/index-ja.html，(参照 2014-7-11)
[8] Chasen 形態素解析器，奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座（松本研究室）