社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2009-ICS-154 (20)
2009 / 3 / 3

概念ベースとEarth Mover's Distanceを用いた文書検索

藤江 悠五† 渡部 広一† 河岡 司†

† 同志社大学大学院工学研究科 〒610-0321 京都府京田辺市多々羅都谷 1-3
E-mail: †dth0705@doshisha.ac.jp, ††{watabe,kawaoka}@indy.doshisha.ac.jp

あらまし 近年，コンピュータとネットワークの発達に伴って，個人が扱える情報は膨大なものとなり，その膨大な情報の中から必要な情報を探し出すのは非常に困難となっている．既存の検索システムは基本的には表記のみを活用するため，意味的には同じ内容の検索でもユーザが入力する語によって検索結果が異なってしまう．そのためユーザが適切なキーワードを考えなければならない．そこで本稿では文書の意味を捉えた検索を実現するために単語の関連性にもとづいた文書間の類似性の定量化手法を提案する．具体的には概念ベースを用い単語間の関連性を求め，Earth Mover's Distanceにより文書間の類似度を計算する方法を提案する．また概念ベースに存在しない固有名詞や新語に対して，Web情報をもとに新概念として意味を定義し，概念ベースを自動的に拡張する方法を提案する．これら提案手法をNTCIR3-WEBによって他の手法と比較実験したところ，本手法が他手法に比べ良好な結果が得られた．
キーワード 情報検索，概念ベース，Earth Mover's Distance, NTCIR

Associative Document Retrieval Using Concept-Base and Earth Mover's Distance

Yugo FUJIE†, Hirokazu WATABE†, and Tsukasa KAWAOKA†

† Department of Knowledge Engineering and Computer Sciences, Doshisha University1-3 Miyakodani, Tatara, Kyotanabe, Kyoto 610-0321, Japan
E-mail: †dth0705@doshisha.ac.jp, ††{watabe,kawaoka}@indy.doshisha.ac.jp

Abstract Recently the development of computers and networks makes amount of information huge. It is very difficult to find necessary information in the huge information.The existing retrieval system uses not the meaning of input words but the notation of them.Therefore, different words bring a defferent result of retreieval even if they have the same meaning.A user of the system has to consider the input words to search the necessary information.This paper proposes the quantification technique of the semantic distance between documents based on relevance of the word to realize the search that captured the meaning of the document.Concretely the related degree between words is calculated by concept-base and the resemblance degree between documents is calculated by Earth Mover's Distance.Besides this paper proposes method that no existence word on concept-base is defined as a concept based on Web information to expand concept-base automatically.Retrieval experiments using the NTCIR3-WEB in comparison with the other method have shown that our method is effective than other method.
Key words Information Retrieval, Concept-Base, Earth Mover's Distance, NTCIR

1. はじめに

一般家庭にもPC，ブロードバンドが普及し，ユーザは手軽に情報を収集できるようになってきている．しかし一方では，情報が過度に溢れ過ぎ，利用者の要求に合った情報を探し出す必要性が高まっている．その中で要求に適合した情報のみを選出するのではなく，情報をランキング付けして提示することも重要となっている．ランキング付けは，検索要求と検索対象との間の類似性や関連性をもとに行われ，これらを定量化することが求められる．その際，従来の情報検索でよく用いられているベクトル空間モデル[1]などでは文書における単語の出現頻度や統計情報などを利用して検索要求と文書間の類似性を判断し，文書を選別している．このような手法は検索要求と文書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もとづいている．しかし，実際の文書において，語の表記が同じでも異なる意味を有したり（多義性），同じ意味でも語の表記

が異なる場合（表記揺れ，同類義語）がある．さらに単語間には，互いに意味的な関連性を持って存在しており，表記だけを頼りに検索を行う手法ではユーザが入力する語によって検索結果が異なってしまう．そのためユーザが適切なキーワードを考えなければならない．その問題を解消するために，ユーザが入力したキーワードの意味を捉えた検索手法が必要である．

このような背景から，本研究では文書における意味を捉えた検索を実現すべく，単語の意味特徴を定義した概念ベース [2] を用いた検索手法を提案する．概念ベースを用いることによって，単語の表記のみでの検索方式とは異なり，意味を捉えた検索が可能になる．つまり，ユーザの入力語の表記的揺らぎに影響されず，検索要求と検索対象間の意味的近さを定量化できる手法である．具体的には，概念ベースによって単語間の意味的な関連性を 0 から 1 までの数値として算出する．そして，その値をもとに検索要求と検索対象との類似度を画像検索等の分野で注目されている距離尺度である Earth Mover's Distance(EMD) [3] により求める方法を提案する．また，概念ベースに存在しない固有名詞や新語に対して，Web をもとに新概念として定義し概念ベースを自動的に拡張する手法を提案する．

2. 先行研究と本研究の位置付け

体系的に整理された辞書である WordNet [4] を用いて単語間の距離を定義し，EMD により文書間の類似度を定義する手法 [5] が提案されている．これにより単語の意味的な関連性に着目した情報検索が実現されている．また，単語の共起情報をもとに単語間の関連性を定義し，EMD により文書間の類似度を定義する手法が提案されている [6]．この手法では，単語の共起情報を用いることにより，全ての単語間の関連性を定義し，文書間の類似度を定義することを実現している．しかしながらこれらの手法の問題点として，WordNet などの整理された辞書を用いる場合は，索引語の全てが辞書に含まれる保証がなく，全ての索引語間の関連性を求めることができない可能性がある．共起情報を手がかりにした場合は，用いる文書集合の特性や容量の影響を大きく受け，正確に関連性を定義しているとは言い難い．

提案手法では，概念ベースを用いて索引語間の関連性を求め，さらに概念ベースに存在しない語においては Web をもとに自動的に概念として定義する．これにより，単語間の関連性をより正確に定義し，さらにあらゆる新語に対応できる索引語の網羅性を実現する．

3. 基本事項

3.1 索引語の取得

本研究は日本語での検索を想定している．日本語は英語などとは異なり，単語間に明確な区切りがない．そこで，文章から単語を切り出す必要がある．本研究では単語の切り出しに茶筌(注1)を用い，「名詞」，「形容詞」，「動詞」を索引語として用いる．

3.2 tf·idf 重み付け

索引語に対する重み付けは，情報検索の分野で広く用いられている tf・idf 重み付け [7] を使用する．tf・idf による重み付けとは，対象としている単語の頻度と網羅性に基づいた重み付け手法である．文書 d における索引語 t の重み $wd(t,d)$ は以下の式 1 によって得られる．

$$wd(t,d) = tf(t,d) \times idf(t) \tag{1}$$

$tf(t,d)$ は文書 d における索引語 t の出現頻度である．ただし，$tf(t,d)$ は文書長の影響を受けやすいため，本論文では以下の式 2 に示す正規化手法を用いた．単語 t の出現頻度を $tfreq(t,d)$，文書 d に含まれる単語数を $tnum(d)$ とする．

$$tf(t,d) = \frac{\log(tfreq(t,d)+1)}{\log(tnum(d))} \tag{2}$$

また，$idf(t)$ は文書数 N と索引語 t が出現する文書の数 $df(t)$ によって決まり，式 3 によって定義される．

$$idf(t) = \log\frac{N}{df(t)} + 1 \tag{3}$$

4. 概念ベース

概念ベースとは複数の国語辞書や新聞などから機械的に構築した単語（概念）とその意味特徴を表す単語（属性）の集合からなる知識ベースである．概念には属性とその重要性を表す重みが付与されている．概念ベースには約 12 万語の概念表記が収録されており，1 つの概念に平均約 30 個の属性が存在する．ある概念 A は属性 a_i とその重み wc_i の対の集合として，式 4 で表される．

$$A = \{(a_1, wc_1), (a_2, wc_2), \cdots, (a_i, wc_i), \cdots, (a_n, wc_n)\} \tag{4}$$

任意の一次属性 a_i は，その概念ベース中の概念表記の集合に含まれている単語で構成されている．したがって，一次属性は必ずある概念表記に一致するため，さらにその一次属性を抽出することができる．これを二次属性と呼ぶ．概念ベースにおいて，「概念」は n 次までの属性の連鎖集合により定義されている（図 1）．

概念	属性／重み
雪	(雪／0.61)，(雪掻き／0.30)，(粉雪／0.27)，…
雪掻き	(雪掻き／0.16)，(除雪／0.14)，(降雪／0.14)，…
粉雪	(粉雪／0.23)，(真っ白／0.21)，(氷点下／0.20)，…

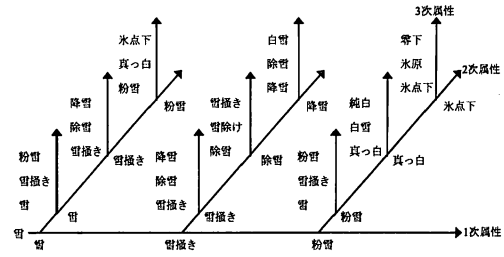

図 1 概念ベース

(注1)：奈良先端科学技術大学院大学．http://chasen-legacy.sourceforge.jp/

5. 概念ベースを用いた単語間の関連性の定量化

概念ベースを用いた単語間の関連性の定量化は，基本的に語意の展開結果を利用し数値として表す．何次属性まで展開するか，どの属性を用いるかによって値が変わってくるため，状況に応じてどのように計算するかが問題になってくる．そこで本研究では二種類の方法を使い分ける．文書間の類似度を求めるための単語間の関連性の定量化には，概念ベースの一次属性までを使用する一致度計算を用い，概念ベースの自動拡張手法における単語間の関連性の定量化には，概念ベースの二次属性までを使用する関連度計算 [8] を使用する．二次属性までを使用する方法が，概念ベースを用いた単語間の関連性の定量化には一番有効であると報告されている [8]．一次属性までしか展開しないと，関連が薄い概念同士の関連性を定量化できず，三次属性まで用いると概念とはかけ離れた語が属性となり，雑音として働くため精度が低下してしまう．本研究では，文書を概念と文書間の類似度を求めるための単語間の関連性の定量化には一次属性までしか展開しない一致度計算を用いる．これは文書を概念と見立てた場合，索引語が一次属性となり，索引語の属性が二次属性となる．つまり，索引語の二次属性まで展開すると文書を概念とした場合の三次属性まで展開したこととなり雑音が増加し，概念（文書）とはかけ離れた語が計算に使用されてしまうためである．

5.1 一致度計算

任意の概念 A，B について，それぞれ一次属性を a_i，b_j とし，対応する重みを u_i，v_j とする．また，概念 A，B の属性数を L 個，M 個 $(L \leqq M)$ とする．

$$A = \{(a_i, u_i) \mid i = 1 \sim L\}$$
$$B = \{(b_j, v_j) \mid j = 1 \sim M\}$$

このとき，概念 A，B の一致度 $MatchWR(A,B)$ を以下の式 5，6 で定義する．

$$MatchWR(A,B) = \sum_{a_i = b_j} \min(u_i, v_j) \tag{5}$$

$$\min(\alpha, \beta) = \begin{cases} \alpha & (\beta > \alpha) \\ \beta & (\alpha \geqq \beta) \end{cases} \tag{6}$$

ただし，各概念の重みの総和をそれぞれ 1 に正規化する．概念 A，B の属性 a_i，b_j に対し，$a_i = b_j$（概念 A，B に共通する属性がある）となる属性があった場合，共通する属性の重みの共通部分，つまり，重みの小さい分だけ有効に一致すると考え，その合計を一致度とする．定義から明らかなように両概念の属性と重みの両方が完全に一致する場合には一致度は 1.0 となる．

5.2 関連度計算

関連度計算は概念の二次属性間の一致度計算により求めた値をもとに概念間の関連性を数値として算出する．具体的には，計算する二つの概念の内，一次属性の数の少ない方の概念を A とし $(L \leqq M)$，概念 A の一次属性を基準とする．

$$A = \{(a_1, u_1), (a_2, u_2), \cdots, (a_i, u_i), \cdots, (a_L, u_L)\}$$

そして概念 B の一次属性を，概念 A の各一次属性との一致度 $MatchWR(a_i, b_{xi})$ の和が最大になるように並び替える．

$$B_x = \{(b_{x1}, v_{x1}), (b_{x2}, v_{x2}), \ldots, (b_{xi}, v_{xi}), \cdots, (b_{xL}, v_{xL})\}$$

これによって，概念 A の一次属性と概念 B の一次属性の対応する組を決める．対応にあふれた概念 B の一次属性は無視する（この時点では組み合わせは L 個）．ただし，一次属性同士が一致する（概念表記が同じ）ものがある場合（$a_i = b_j$）は，別扱いにする．これは概念ベースには約 12 万の概念表記が存在し，属性が一致することは稀であるという考えに基づく．従って，属性の一致の扱いを別にすることにより，属性が一致した場合を大きく評価する．具体的には，対応する属性の重み u_i，v_j の大きさを重みの小さい方にそろえる．このとき，重みの大きい方はその値から小さい方の重みを引き，もう一度，他の属性と対応をとることにする．例えば，$a_i = b_j$ で $u_i = v_j + \alpha$ とすれば，対応が決定するのは (a_i, v_j) と (b_j, v_j) であり，(a_i, α) はもう一度他の属性と対応させる．このように対応を決めて，対応の取れた属性の組み合わせ数を T 個とする．このとき，概念 A，B の一致度 $DoA(A,B)$ を以下の式 7 により定義する．

$$DoA(A,B) = \sum_{i=1}^{T} \{MatchWR(a_i, b_{xi}) \times (u_i + v_{xi}) \times (\min(u_i, v_{xi}) / \max(u_i, v_{xi}))/2\} \tag{7}$$

関連度の値は概念間の関連の強さを 0〜1 の間の連続値で表す．1 に近づくほど関連が強い．本研究では 6.3 節で述べる概念ベースに属性として追加する概念と追加される側の概念との間の関連性の定量化に関連度計算を用いる．

6. 概念ベースの自動拡張手法

概念ベースに存在しない語（未定義語）にも属性を与えなければ，未定義語と他の単語との関連性を求めることができない．そこで，現在考えうる最大の言語データである Web 情報をもとに未定義語の概念化を行い，概念ベースに追加する方法を提案する．本章ではこの手法について説明する．

6.1 未定義語の概念化

未定義語の概念化を行うために，未定義語の属性とその重みを Web から以下の手順で獲得する．まず，未定義語をキーワードとして検索エンジンを用いて検索を行い，検索上位 100 件の検索結果ページの内容を取得する．次に，取得した文書群に対して，茶筌を用いて形態素解析を行い，自立語を抽出する．最後に，概念ベースに存在する語のみを未定義語の属性とし，頻度情報と特定性情報である idf 値を掛け合わせたものを属性の重みとする．

6.2 属性内出現頻度を用いた重み付け手法

概念ベースに未定義語を追加する場合には概念ベースの頻度情報を用いて重み付けを行う必要がある．概念に付与された属性は，特徴を表す語であるため，概念の説明文書であると捉えることが出来る．この文書空間内での属性の出現頻度を，概念に対する属性の確からしさだと考える．例えば個人情報という概念を特徴付ける一次属性には，「個人，情報，識別，…」という属性が存在する．これは，「個人を識別することができる情報

を指す」という文書であると捉えることができる．このように，概念に対する n 次属性空間はその概念についての説明文書の集合だとみなすことができる．この n 次属性空間から算出した出現頻度を n 次属性内出現頻度と呼ぶ．本研究では 2 次属性内出現頻度を用いる．3 次属性空間までを用いると，概念に関係のない語が多くなってしまうためである．

未定義語の属性 A の 2 次属性内出現頻度を $freq(A)$，未定義語の一次属性の総数を R とし，未定義語の属性 A の概念ベース空間の idf 値を $cidf(A)$ とすると，重み $wc(A)$ は以下の式で表される．

$$wc(A) = \frac{\log(freq(A))}{\log(R)} \cdot cidf(A) \tag{8}$$

6.3 相互追加

新規概念に属性を追加した場合，その概念自身は属性として他の概念を持つが，その概念を属性として持つ概念は存在しない．したがって，他の概念に新しく追加をした概念を属性として追加する手法が必要となる．新概念とその獲得した属性にはWeb 上のホームページにともに出現しており，共起という関係があり，属性から見ても新概念とのなんらかの関連性があると考えられる．このため，新概念を属性の追加候補とする．しかし，全ての属性追加候補を属性として追加すると雑音が非常に多くなってしまうため，選別して追加を行う．新概念から取得した語に対して新概念を属性として選別して追加することを相互追加と呼ぶこととする．

追加する概念 A と追加される概念 B との間の関連度 $DoA(A,B)$ と追加する属性の概念ベース空間の idf 値を $cidf(A)$ とすると追加した属性の重み $wc(A)$ は以下の式で表される．

$$wc(A) = DoA(A,B) \cdot cidf(A) \tag{9}$$

7. EMD を用いた文書検索

検索要求と検索対象の間の類似度を求める際，いくら単語間の関連性を正確に定義できたとしても，その値をもとにうまく計算できないと文書間の正確な類似度を求めることはできない．計算の仕方としては，様々な方法が考えられ，例えば単語間の関連性が高い順に単語の対応をとり計算する方法などが挙げられる．1 対 1 で対応をとる方法では，検索要求と検索対象の語の少ない方の語の数しか対応がとれない．例えば，検索要求の語が 3 語，検索対象の語が 100 語であった場合，検索対象の 97 語は計算の対象外となる．さらに，実際の検索において，ユーザは検索要求にあまり多くの語を入力しないと考えられ，検索要求と検索対象との語数の差は非常に大きいと想定され，文書内の単語の重要性と単語間の関連性を考慮し M 対 N で柔軟に対応を取る必要がある．

そこで本研究では類似画像検索の分野で注目されている EMD [3] を用いて文書間の類似度を算出する方法を用いる．EMD は輸送問題における輸送コストの最適解を求めるアルゴリズムであり，需要地 (供給地) の重みと需要地と供給地間の距離を定義できればどのような問題にも適用できる．この EMD を用いることで単語の重みと単語間の関連性を考慮して柔軟に対応を取り，文書間の類似度を求めることができる．

7.1 EMD とは

EMD は線形計画問題の一つであるヒッチコック型輸送問題において計算される距離尺度であり，2 つの離散分布において，一方の分布を他方の分布に変換するための最小コストとして定義される．輸送問題とは，需要地の需要を満たすように供給地から需要地へ輸送を行う場合の最小輸送コストを解く問題である．

EMD を求める際，二つの分布は要素の重み付き集合として表現される．一方の分布 P を集合として表現すると，$P = \{(p_1, w_{p_1}), ..., (p_m, w_{p_m})\}$ となる．今，分布 P は m 個の特徴量で表現されており，p_i は特徴量，w_{pi} はその特徴量に対する重みである．同様に，一方の分布 Q も集合として表すと，$Q = \{(q_1, w_{q_1}), ..., (q_n, w_{q_n})\}$ となる．EMD の計算は，2 つの分布において特徴量の数が異なっている場合でも計算が可能であるという性質を持っている．今，p_i と q_j の距離を d_{ij} とし，全特徴間の距離を $D = [d_{ij}]$ とする．ここで，p_i から q_j への輸送量を f_{ij} とすると，全輸送量は $F = [f_{ij}]$ となる．ここで，式 10 に示すコスト関数を最小とする輸送量 F を求め，EMD を計算する．

$$WORK(P,Q,F) = \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij} \tag{10}$$

ただし，上記のコスト関数を最小化する際，以下の制約条件を満たす必要がある．

$$f_{ij} \geqq 0, \quad 1 \leqq i \leqq m, \quad 1 \leqq j \leqq n \tag{11}$$

$$\sum_{j=1}^{n} f_{ij} \leqq w_{p_i}, \quad 1 \leqq i \leqq m \tag{12}$$

$$\sum_{i=1}^{m} f_{ij} \leqq w_{q_j}, \quad 1 \leqq j \leqq n \tag{13}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}\right) \tag{14}$$

ここで，式 11 は輸送量が正であることを表し，p_i から q_j に送られる一方通行であることを表している．式 12 は輸送元である p_i の重み以上に輸送できないことを表す．式 13 は輸送先である q_j の重み以上に受け入れることができないことを表す．最後に式 14 は総輸送量の上限を表し，それは輸送先または輸送元の総和の小さい方に制限されることを表す．以上の制約条件の下で求められた最適な全輸送量 F を用いて分布 P，Q 間の EMD を以下のように求める．

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \tag{15}$$

ここで，最適なコスト関数 $WORK(P,Q,F)$ を EMD としてそのまま用いないのは，コスト関数は輸送元もしくは輸送先の重みの総和に依存するので，正規化することによってその影響を取り除くためである．

8. EMDの文書検索への適用

図2にEMDを文書検索に適用した例を示す.

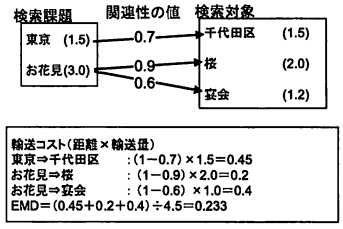

図2 EMDを文書検索に適用した例

EMDを文書検索に適用する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．需要地としては，検索課題の索引語を，供給地としては検索対象の索引語を割り当てる．需要量と供給量はそれぞれ索引語の3.2節で説明したtf・idf重みを用いる．そして，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，提案手法においては概念ベースを用いた一致度計算により求めることができる．一致度は関連性が高いと値も大きくなるため，1から一致度の値を引いた値に変換する．EMDの計算は図2の下方で求まる．「お花見」と「宴会」間の輸送量が1となっているのは，「お花見」から「桜」に重み2を輸送し，「お花見」の余った重み1を「宴会」に輸送したためである．このように，関連性が高い語に優先して重みを輸送し，供給量がなくなるか需要量が満たされるまで輸送を行う．このように索引語間の関連性と重みを考慮したM対Nでの柔軟な対応が可能である．EMDの特徴として，索引語間の距離の値が0から1であるなら，EMDも0から1の値になる．そして，EMDは文書間が似ていると値が低くなり，似ていないと値が高くなる．よって値が低い文書から順にユーザに提示することで文書検索が実現できる．

9. 実験と評価

単語の関連性に着目した提案手法の有効性を検証するため情報検索システムテストコレクションNTCIR3-WEB(注2)を用いて，表記を用いる他の手法との比較実験を行った．比較手法としては，ベクトル空間モデル[1]，Okapi BM25[9]と，同じ索引語間の距離を0，それ以外を1とした素朴なEMDを用いた．

9.1 評価方法

今回の評価では，検索課題41件と正解文書とランダムに選択した文書を合わせた10,000件の文書を使用し，評価実験を行った．また，正解文書リストが存在し，各検索課題に対して，各文書がH（高適合），A（適合），B（部分的適合），C（不適合）の4段階の適合度が設定されているが，今回はHとAを正解とした．

各検索課題に対して，10000件の検索対象全てのスコアを求め，スコア順に並べ変える．そして，正解文書リストを参照し正解文書の順位を調べ評価する．

9.2 評価指標

評価指標には，各検索課題毎の平均精度（Avelage Precision, AP），平均精度の平均（Mean Average Precision, MAP）と再現率ー精度グラフを使用した．検索課題に対する平均精度APは式16のように定義される．まず順位i位の文書が適合しているならば1，そうでなければ0となる変数をz_iとする．Sを適合文書の総数，nは出力文書数である．

$$AP = \frac{1}{S}\sum_{i=1}^{n}\frac{z_i}{i}\left(1+\sum_{k=1}^{i-1}z_k\right) \tag{16}$$

平均精度の平均(MAP)は，全ての検索課題に対して平均精度を平均したものであり，式17によって求められる．具体的には，検索課題がK件ありそれぞれの課題に対するあるシステムの平均精度をAP_hと表記すれば$(h=1,...,K)$，その平均がMAPに相当し，以下の式に示す．

$$MAP = \frac{1}{K}\sum_{h=1}^{K}AP_h \tag{17}$$

再現率ー精度グラフとは再現率の11個の点ごとに，41個の検索課題の精度を平均してグラフを描いたものである．

9.3 比較手法

本節では，提案手法との比較に用いているベクトル空間モデルとOkapiBM25について述べる．

9.3.1 ベクトル空間モデル

ベクトル空間モデルは，情報検索の分野で幅広く利用されている検索モデルである．各語の重みから構成されるベクトルとして検索課題と文書をそれぞれ表現し，二つのベクトルの成す角度の余弦によって類似度を計算する点に特徴がある．重みの種類にはいくつかの種類があるが，本実験では3.2節で説明したtf・idf重みを用いる．検索課題qと文書d_iの索引語の語の総数（異なり）をMとすれば，文書と検索課題はそれぞれ以下のようなM次元ベクトルで表現できる．

$$d_i = (w_{i1}, w_{i2}, \cdots, w_{iM}) \tag{18}$$

$$q = (w_{q1}, w_{q2}, \cdots, w_{qM}) \tag{19}$$

検索課題qに対する文書d_iの得点$s_q(d_i)$は2つのベクトルの角度の余弦により求まる．式を以下に示す．

$$s_q(d_i) = \frac{\sum_{j=1}^{M} w_{ij}w_{qj}}{\sqrt{\sum_{j=1}^{M} w_{ij}^2}\sqrt{\sum_{j=1}^{M} w_{qj}^2}} \tag{20}$$

9.3.2 Okapi BM25

S.E.Robertsonを中心に開発されたOkapiと呼ばれる次世代検索システムにおいて使用されている確率型の検索モデルBM25は，ベクトル空間モデルと同等，あるいはそれ以上の性能を示すことでよく知られている．原理的には，検索課題qと

(注2)：国立情報学研究所.
http://research.nii.ac.jp/ntcir/ntcir-ws3/ws-ja.html

文書ベクトル d_i が与えられた時に，その文書が検索課題に適合している確率を推計するものである．検索課題 q と文書 d_i の索引語の語の総数（異なり）を M とすれば，検索課題 q に対する文書 d_i の得点 $s_q(d_i)$ は以下の式で表される．

$$s_q(d_i) = \sum_{j=1}^{M} (w_{ij} \times x_{qj} \times \tau_j) \tag{21}$$

ただし，x_{qj} は検索課題 q での語 t_j の出現回数である．ここで

$$w_{ij} = \frac{3.0 x_{ij}}{(0.5 + 1.5 l_i/\bar{l}) + x_{ij}} \tag{22}$$

$$\tau_j = \log \frac{N - n_j + 0.5}{n_j + 0.5} \tag{23}$$

である．x_{ij} は文書 d_i での t_j の出現回数であり，N は文書総数で，n_j は語 t_j が出現する文書数である．なお，

$$l_i = \sum_{j=1}^{M} x_{ij} \tag{24}$$

は文書 d_i の長さであり，

$$\bar{l} = \frac{1}{N} \sum_{i=1}^{N} l_i \tag{25}$$

はデータベース全体での文書長の平均を意味する．

9.4 評価結果

MAP は提案手法では 0.5214，ベクトル空間モデルでは 0.4334，Okapi BM25 では 0.4422，素朴な EMD では 0.4774 であった．提案手法の精度はベクトル空間モデルより 20.30%，Okapi BM25 より 17.91%，EMD より 9.22%の精度向上を達成した．再現率ー精度グラフを図 3 に示す．

図 3 より，全ての再現率レベルで精度が改善している．

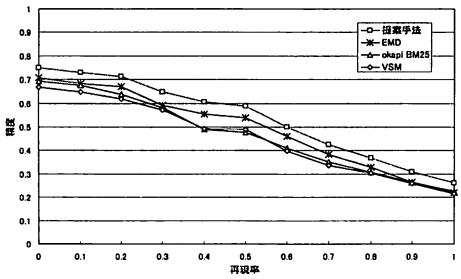

図 3 再現率ー精度グラフ（提案手法と表記に頼る手法との比較）

10. 概念ベースの自動拡張手法の評価

概念ベースの自動拡張手法の効果を検証するために，自動拡張手法を用いない手法（概念ベースに存在しない索引語で同じ索引語間の距離を 0 それ以外の距離を 1）と提案手法との比較評価を行った．評価方法は 9.1 節と同様の方法で行った．以下，自動拡張手法を用いない EMD と概念ベースを組み合わせた手法を EMD + CB と記す．MAP は提案手法では 0.5214，EMD+CB では 0.5133 であった．再現率ー精度グラフを図 4 に示す．

図 4 より差は小さいが全ての再現率レベルで提案手法が EMD+CB を上回っている．

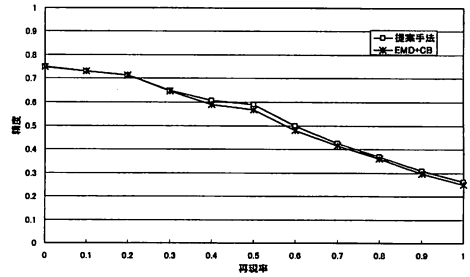

図 4 再現率ー精度グラフ (自動拡張手法使用と未使用での比較)

11. おわりに

本論文では，索引語の関連性を概念ベースにより定義し，それをもとに EMD によって文書間の類似性を求める手法を提案した．さらに概念ベースに存在しない語においては Web をもとに語の意味を定義し概念ベースを自動的に拡張することで対応し，全ての索引語間の距離を概念ベースにより求めることを可能とする手法を提案した．そして，その有効性を Web 検索評価用テストコレクション NTCIR3-WEB を用いて検証した．

結果として，表記に頼る他の手法に比べ良好な結果を得て，単語の関連性に着目した本手法の有効性を確認できた．また Web により概念ベースを自動的に拡張する手法を使用と未使用で比較評価を行い，Web により未定義語を定義することの有効性を示し，あらゆる新語に対応できる索引語の網羅性を実現した．今後は情報検索に限らず文書分類など様々な分野へ応用していきたい．

文　献

[1] G.SaltonA.WongC.S.Yang, A Vector space model for automatic indexing, Communications of the ACM, pp.613–6201, Vol.18, No.3, 1975.

[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] Y.Rubner, C.Tomasi, L.Guibas, The earth mover's distance as a metric for image retrieval, Int.J.Comput.Vision, pp.99–121, vol.40, 2000.

[4] G.A.Miller,WordNet: A lexical database for English,Commun. ACM, pp.39–41, Vol.38, No.11, 1995.

[5] X.Wan, Y.Peng, The Earth Mover's Distace as a Semantic Measure for Document Similality, Proceeding of the 14th ACM international conference on Information and knowledge management, pp.301–302, 2006.

[6] 柳本泰一, 大松繁, Earth Mover's Distance を類似度として用いた情報検索, 電気学会全国大会, 3-065, 2007.

[7] G.Salton, C.Buckley, Term-weighting approaches in automatic text retrieval, Information Processing and Management, pp.513–523, Vol.41, No.4, 1988.

[8] 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,pp53–74, Vol.13,No.1, 2006.

[9] S.E.Robertson, S.Walker, S.Jones and M.Beaulieu, M.Gatford, Okapi at TREC-3, In proceeding of the 3rd Text Retrieval Conference,pp109–126,1995