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[†]

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[†]

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*

[†] 同志社大学大学院工学研究科知識工学専攻, Dept. of Knowledge Engineering & Computer Sciences, Graduate School of Engineering Doshisha University

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 (Salton, Wong, and Yang 1975) 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 (奥村, 土屋, 渡部, 河岡 2007) 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, 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) (Rubner, Tomasi, and Guibas 2000) 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 (Miller 1995) 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 (Wan and Peng 2006) 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 (柳本, 大松 2007). 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 NTCIR3-WEB

本研究では提案手法の効果を検証するため, 日本での代表的な情報検索システム評価用のテストコレクションである NTCIR¹を用いた. NTCIR は文献データ集合, 検索課題集合, 各検索課題に対する文献の適合不適合判定からなるもので, 同一のテストコレクションを利用することにより共通の基準で情報検索システムを評価することができるようにしたものである. その中でも, 本研究では一般の利用者が実際に検索する環境に近い Web 検索用のテストコレクションである NTCIR3-WEB を用いた. 図 1 に NTCIR3-WEB の検索課題の一例を示す. 検索課題には NUM, TITLE, DESC, NARR, CONC, RDOC, USER の 7 つのフィールドが含まれているが, このうち標準的なフィールドは TITLE (title), DESC (description), NARR (narrative), CONC (concept) の 4 つである. TITLE とは検索課題の内容を簡単に表したタイトル, DESC は検索する内容を文で記述したもの, NARR は検索する内容の詳細な説明, CONC は検索する内容を表すキーワードである. 本研究では検索要求を文章で入力するシステムの開発を想定し,

```
<TOPIC>
<NUM>0008</NUM>
<TITLE CASE="b">サルサ, 学ぶ, 方法</TITLE>
<DESC>サルサを踊れるようになる方法が知りたい</DESC>
<NARR><BACK>最近はやっているサルサという踊りを学ぶためにどうすればよいか具体的な方法が知りたい. 例えば教室に通うという場合には, その場所や授業形態など, 具体的な内容を必要とする.</BACK><RELE>具体的な方法の表記のない, 流行であることのみを扱った文書は不適合とする.</RELE></NARR>
<CONC>サルサ, 習う, 方法, 場所, カリキュラム</CONC>
<RDOC>NW011992774, NW011992731, NW011992734</RDOC>
<USER>大学院修士 1 年, 女性, 検索歴 2.5 年</USER>
</TOPIC>
```

図 1 NTCIR3-WEB の検索課題の例

¹ 国立情報学研究所. <http://research.nii.ac.jp/ntcir/ntcir-ws3/ws-ja.html>

SalsaLinksFireworksSpliceHTMLイベント情報やお知らせなどを、随時更新しています。詳しい情報や裏情報をいち早くGETしたい人は、チェックをお忘れなく！ダンス・スクールからのお知らせも、こちらからリンクしています。SHJからのお知らせやイベント情報の詳細を、随時更新。まずは、こちらをチェック！情報責任者：ジョージ渡部
 2001年11月10日・11日に開催が決定した、JapanSalsaCongressの情報を随時更新。/情報責任者：ジョージ渡部SHJのHP内でご紹介しているダンススクールからの、イベントやレッスンのお知らせです。（この欄やスクール紹介のページに掲載ご希望のインストラクターの方は、SHJまでご連絡をお願い致します。）
 ●DANZALATINA/情報責任者：中野智史
 ●LOSSALSEROS/情報責任者：石川裕子
 ●ZERO/情報責任者：YUKI
 ●SalsaVenom/情報責任者：竹内雅彦
 全国のサルサイバントやライブなど、主催者や出演者からのお知らせのページです。

図 2 NTCIR3-WEB の検索対象の例

DESCのみを使用する。図2に検索対象（HTMLもしくはプレーンテキストファイル、言語は主に日本語と英語、ごく一部にその他の言語、100 GB）の一例を示す。実験には検索対象のタグを省いた全文を用いる。

3.2 索引語の取得と重み付け

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。本研究では形態素解析器を用いて行う。

3.2.1 形態素解析器

日本語構造の制約を利用し、単語の切り出しや品詞を同定することを形態素解析という。例えば、形容詞は名詞の前に付くことができるという法則である。実際は、日本語の複雑さのため完全に単語を切り出すことは難しいが、代表的な形態素解析器である茶筌²は様々な工夫により高い精度で単語を正しく切り出すことができる。本研究では単語の切り出しに茶筌を用い、「名詞」、「形容詞」、「動詞」を索引語として用いる。

3.2.2 tf・idf

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 (Salton and Buckley 1988) 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。

$$wd(t, d) = tf(t, d) \times idf(t) \quad (1)$$

² 奈良先端科学技術大学院大学。http://chasen-legacy.sourceforge.jp/

索引語	検索課題 idf
:	:
映画	4.554589
どこ	4.554589
使う	4.554589
作品	4.554589
種類	4.554589
含む	4.554589
的	3.969626
様々	3.969626
なる	3.554589
日本	3.554589
効果	3.554589
書く	3.554589
もの	3.232661
よう	2.969626
:	:

削除

図 3 検索課題 idf による不要語の削除

$tf(t, d)$ 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。

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

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。

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

3.2.3 不要語の削除

本研究では不要語を削除するために評価データの実験に使用する検索対象の空間での idf 値をもとに検索対象内の不要語を削除した。不要語とは「する」や「こと」のような, どの文書にも出現し, 文書を特定するために有効でない語を指す。検索対象の不要語削除の閾値は提案手法が適合判定レベルの LEVEL1 においての評価 (MAP) が idf 値 0 から 9 の間で一番高くなる idf 値 5 に設定した。適合判定 LEVEL と評価 (MAP) などの評価方法については 8.1 節で述べる。さらに検索課題空間での idf 値を利用し検索課題の索引語中の不要語を削除した。削除の例を図 3 に示す。これにより検索課題によく出現する「文書」や「様々」などの語を削除でき, 検索課題として比較的意义があると考えられる「将来」, 「歴史」, 「地域」などを残すことができる。この閾値は目視により設定した。

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．

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

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 4）．以下に，本研究で使用した大規模概念ベースの構築方法について述べる．

4.1 概念ベースの構築方法

まず，基本となる概念ベースを複数の国語辞書から構築する．概念は国語辞書の見出し語から，属性は見出し語の語義説明文の自立語から，その重みは自立語の出現頻度に基づいて決定される（笠原，松澤，石川 1997）．そして，属性信頼度（語に関する種々の知識から属性として

図 4 概念ベース

の確からしさを定量化した値)により不適切な属性の削除を行い, 信頼性の高い属性を抽出する(小島, 渡部, 河岡 2002). これらにより概念数約 3 万 4 千語で平均属性数が 16 個の基本となる概念ベースが構築される.

複数の国語辞書には約 12 万語の見出し語(辞書に収録されている約 20 万語語彙のうち見出し語として不適切な表記を除去した語)があり, 辞書中の語義説明文だけでは属性が付与できない語が約 9 万語もあった. そこで, 概念総数と属性数を拡張するために電子化新聞(毎日新聞, 日本経済新聞)を用いて, 各概念に対する共起語を属性候補として追加する. この際, もともと概念ベースに定義されている概念についても, 同様に属性を取得する. その後, 5.2 節で説明する関連度計算により概念と属性の関連の強さを求め, その値と概念ベースの頻度情報をもとに属性の重み付けを行う. 以上の処理により本研究で使用した概念数が約 12 万語, 平均属性数約 30 個の大規模概念ベース(奥村他 2007)が構築できる.

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(渡部, 奥村, 河岡 2006)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(渡部他 2006). 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 ($L \leq M$) 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.

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

$$\min(\alpha, \beta) = \begin{cases} \alpha & (\beta > \alpha) \\ \beta & (\alpha \geq \beta) \end{cases}. \quad (6)$$

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 ($L \leq M$), 概念 A の一次属性を基準とする.

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

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.

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

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.

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

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．概念 A と B に対して関連度計算を行った例を表 1 に挙げる．最後に，概念「机」と「椅子」を例に用いて，関連度の計算例を説明する．概念「机」と「椅子」の一次属性および二次属性を表 2，表 3 に示す．

まず，概念「机」と「椅子」の一致度の計算を行う．例えば，概念「机」の一次属性「学校」と概念「椅子」の一次属性「木」は，「木造」という共通する属性を持っているため，一致度は以下のように計算される．

$$MatchWR(\text{学校, 木}) = \min(0.2, 0.4) = 0.2$$

同様に全ての一次属性の組み合わせについて一致度を計算した結果を表 4 に示す．

次に，関連度の計算を行う．関連度の計算は，まず属性が完全に一致している部分から行われる．続いて，一致度の大きい部分から順に対応を決める．この場合，表 4 から一次属性「勉強」と「勉強」，「学校」と「教室」，「本棚」と「勉強」の順に対応が決まることになる．結果，

表 1 関連度計算の例

概念 A	概念 B	関連度の値
自動車	車	0.912
自動車	学校	0.001
天気	晴れ	0.446
天気	学校	0.002

表 2 概念「机」と「椅子」の一次属性

机		椅子	
属性	重み	属性	重み
学校	0.6	勉強	0.5
勉強	0.3	教室	0.3
本棚	0.1	木	0.2

表 3 概念「机」と「椅子」の二次属性

学校		勉強		本棚		教室		木	
属性	重み	属性	重み	属性	重み	属性	重み	属性	重み
大学	0.4	予習	0.5	図書	0.6	教師	0.4	森林	0.5
校舎	0.4	試験	0.3	書物	0.3	校舎	0.4	木造	0.4
木造	0.2	本	0.2	本	0.1	生徒	0.2	葉	0.1

表 4 概念「机」と「椅子」の一致度行列

	学校	勉強	本棚
勉強	0.0	1.0	0.1
教室	0.4	0.0	0.0
木	0.2	0.0	0.0

関連度は次式のように計算される.

$$\begin{aligned} DoA(\text{机, 椅子}) &= 1.0 \times (0.3 + 0.3) \times (0.3/0.3)/0.2 + 0.4 \times (0.6 + 0.3) \times (0.3/0.6)/2 \\ &\quad + 0.1 \times (0.1 + 0.2) \times (0.1/0.2)/2 \\ &= 0.3975 \end{aligned}$$

本研究では 6.4 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

以下の手順により、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 から獲得する.

- (1) 入力された未定義語をキーワードとして検索エンジン³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.
- (2) HTML タグなど不要な情報を取り除いた文書群に対して、形態素解析を行い、自立語を抽出する.
- (3) 得られた自立語の中から概念ベースに存在する単語のみを未定義語の属性として抜き出す.
- (4) 得られた属性の頻度に Web 上の単語の idf を統計的に調べたもの (SWeb-idf) (辻, 渡部, 河岡 2004) の値を掛け合わせたものを属性の重みとし、得られた重み順に並び替える. SWeb-idf については次節で説明する. なお, SWeb-idf のデータベースに存在しない属性については, Web 上にあまり存在しない単語と考え, SWeb-idf 値の最大値を掛け合わせている.

表 5 に未定義語を概念化した例を示す.

³ 検索エンジンは google を用いた. <http://www.google.co.jp>

表 5 未定義語「G ショック」, 「クイニーアマン」の属性とその重み (一部)

G ショック* ¹		クイニーアマン* ²	
属性	重み	属性	重み
時計	358.94	パン	382.88
腕時計	280.64	バター	249.17
電波	254.72	菓子	116.73
タフ	174.57	生地	90.98
モデル	140.65	ケーキ	82.95

*¹ 「G ショック」は腕時計の一種*² 「クイニーアマン」は菓子パンの一種

6.2 SWeb-idf

SWeb-idf (Statics Web-Inverse Document Frequency) とは, Web 上の単語の idf を統計的に調べた idf 値である. まず, 無作為に選んだ固有名詞 1,000 語を作成する. 表 6 に無作為に選択した固有名詞の一部を示す. この作成した 1,000 語に対して個々に検索エンジンで検索を行い, 1 語につき検索上位 10 件の検索結果ページの内容を取得する. よって, 得られた検索結果ページ数は 10,000 ページとなる. この 10,000 ページから, 複数の国語辞書や新聞などから概念 (単語) を抽出した知識ベースである概念ベースの収録語数である約 12 万語とほぼ同等の単語数が得られたことから, 獲得した 10,000 ページを Web の全情報情報空間とみなしている. そして, その中での単語の idf 値を表す SWeb-idf は, 式 8 で求められる.

$$SWeb-idf(t) = \log \frac{N}{df(t)} \quad (N = 10000) \quad (8)$$

これらにより得られた単語とその idf 値をデータベースに登録した. なお $df(t)$ 項は, 全文書空間 (10,000 ページ) に出現する概念 t のページ数である. 獲得した SWeb-idf の値の例を表 7 に挙げる. なお, 固有名詞の選び方を変えても SWeb-idf の値に大きな変化は見られないという報告がなされている (辻他 2004).

表 6 SWeb-idf の作成に用いた固有名詞 (一部)

固有名詞
舩添要一
SARS
フェラーリ
ラブラドル
筑波技術短期大学

表 7 SWeb-idf の例

概念	SWeb-idf の値
金粉	13.29
焼きたて	13.29
⋮	⋮
株式会社	5.60
⋮	⋮
キーワード	0.00
ページ	0.00

6.3 属性内出現頻度を用いた重み付け手法

未定義語の属性の重みは SWeb-idf によっても求まるが, Web 情報と概念ベースでは語の頻度情報が異なり重みの値も変わるため, Web 情報の重みをそのまま用い概念ベースに追加すると概念ベースの頻度情報に歪みが生じる. よって SWeb-idf は未定義語の属性候補を獲得する時のみ用い, 未定義語の属性の重み付けには使用せず, 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. そこで, 未定義語の属性に対して重みを付与する方法として, 概念ベースの属性空間を考慮した重み付け手法を提案する. 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できる. この文書空間内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. 例として「個人情報」の属性を表 8 に示す. 網掛けのセルに一次属性, 各網掛けの下方のセルにはその二次属性を示している.

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. 本稿では 2 次属性空間を用いる. 3 次属性空間までを用いると, 概念「個人情報」に対して 2 次属性「力」の属性「動物, 筋肉, 物体, …」が含まれる. このため, 概念に関係のない語が多くなってしまうためである.

3.2.2 節で説明した $tf \cdot idf$ 重みの考え方をもとに, 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.

$$wc(A) = \frac{\log(freq(A))}{\log(R)} \cdot cidf(A) \quad (9)$$

6.4 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

表 8 新概念「個人情報」の一次, 二次属性

個人	情報	識別	法律	指す
人	事物	判断	裁判	指
人件費	出来事	色	裁判所	先
擬人	様子	調べる	行政	情報
公私	内容	力	個人	指先

表 9 新概念の属性と重み

概念	属性
ワンセグ	(放送, 0.46), (テレビ, 0.21), (車両, 0.20)

表 10 2 対の関係がある概念の組の例

基本概念	関係がある概念
椅子	机
医師	看護婦
飲料	喉
⋮	⋮

て追加する手法が必要となる。新概念と属性の例を表 9 に示す。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。例であると、概念「放送」、「テレビ」、「車両」に「ワンセグ」という語を属性追加候補とする。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。概念「車両」に対して「ワンセグ」という属性は不必要であるため、選別して追加を行う。このように、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。選別方法としては、2 次属性内出現頻度を属性数で割った値（以下 2 次属性内出現頻度割合と記述）が 0.149 以上かつ関連度 0.068 以上の場合に追加を行う。選別のための閾値の設定は実験により定めた。実験方法としては、概念とその概念と関係がある概念の組を 1,780 組集め、その全ての組における 2 次属性内出現頻度割合と関連度を求め、その平均値を閾値に設定した。実験に使用した 2 対の関係がある概念の組を表 10 に示す。

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。

$$wc(A) = DoA(A, B) \cdot cidf(A) \quad (10)$$

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 で対応を取る方法では、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和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 (Rubner et al. 2000) 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、 $P = \{(p_1, w_{p_1}), \dots, (p_m, w_{p_m})\}$ 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_{p_i} 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、 $Q = \{(q_1, w_{q_1}), \dots, (q_n, w_{q_n})\}$ 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 11 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。

$$WORK(P, Q, F) = \sum_{i=1}^m \sum_{j=1}^n d_{ij} f_{ij} \quad (11)$$

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。

$$f_{ij} \geq 0, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n \quad (12)$$

$$\sum_{j=1}^n f_{ij} \leq w_{p_i}, \quad 1 \leq i \leq m \quad (13)$$

$$\sum_{i=1}^m f_{ij} \leq w_{q_j}, \quad 1 \leq j \leq n \quad (14)$$

$$\sum_{i=1}^m \sum_{j=1}^n f_{ij} = \min \left(\sum_{i=1}^m w_{p_i}, \sum_{j=1}^n w_{q_j} \right) \quad (15)$$

ここで, 式 12 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 13 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 14 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 15 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.

$$EMD(P, Q) = \frac{\sum_{i=1}^m \sum_{j=1}^n d_{ij} f_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}} \quad (16)$$

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.

7.2 EMD の文書検索への適用

図 5 に EMD を文書検索に適用した例を示す. 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 3.2.2 節で説明した $tf \cdot idf$ 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で見立てることができる. 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 5 の下方で求まる. 「梅」と「祭り」間の輸送量が 1 となっているのは, 「梅」から「うめ」に重み 2 を輸送し, 「梅」の余った重み 1 を「祭り」に輸送し

図 5 EMD を文書検索に適用した例

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。

8 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 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、Okapi BM25 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。

8.1 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おり、以下の基準で評価する。

LEVEL1: H 判定と A 判定を適合

LEVEL2: H 判定と A 判定と B 判定を適合

各検索課題に対して、10,000 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。

8.2 評価指標

評価指標には、各検索課題毎の平均精度 (Average Precision, AP)、平均精度の平均 (Mean Average Precision, MAP) と再現率—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 17 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。

$$AP = \frac{1}{S} \sum_{i=1}^n \frac{z_i}{i} \left(1 + \sum_{k=1}^{i-1} z_k \right) \quad (17)$$

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 18 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 ($h = 1, \dots, K$)、その平均が MAP に相当し、以下の式に示す。

$$MAP = \frac{1}{K} \sum_{h=1}^K AP_h \quad (18)$$

再現率—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。

8.3 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 (Salton et al. 1975) と OkapiBM25 (Robertson, Walker, Jones, Beaulieu, and Gatford 1995) について述べる。

8.3.1 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 3.2.2 節で説明した $tf \cdot idf$ 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。

$$d_i = (w_{i1}, w_{i2}, \dots, w_{iM}) \quad (19)$$

$$q = (w_{q1}, w_{q2}, \dots, w_{qM}) \quad (20)$$

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。

$$s_q(d_i) = \frac{\sum_{j=1}^M w_{ij} w_{qj}}{\sqrt{\sum_{j=1}^M w_{ij}^2} \sqrt{\sum_{j=1}^M w_{qj}^2}} \quad (21)$$

8.3.2 Okapi BM25

S. 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。

$$s_q(d_i) = \sum_{j=1}^M (w_{ij} \times x_{qj} \times \tau_j) \quad (22)$$

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

$$w_{ij} = \frac{3.0x_{ij}}{(0.5 + 1.5l_i/\bar{l}) + x_{ij}} \quad (23)$$

$$\tau_j = \log \frac{N - n_j + 0.5}{n_j + 0.5} \quad (24)$$

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,

$$l_i = \sum_{j=1}^M x_{ij} \quad (25)$$

は文書 d_i の長さであり,

$$\bar{l} = \frac{1}{N} \sum_{i=1}^N l_i \quad (26)$$

はデータベース全体での文書長の平均を意味する.

8.4 評価結果

平均精度の平均 (MAP) を表 11 に示す. 再現率—精度グラフを図 6, 7 に示す.

表 11 より, LEVEL1 では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, LEVEL2 では提案手法の精度はベクトル空間モデルより 10.71%, Okapi BM25 より 9.45%, EMD より 4.03%の精度向上を達成した.

図 6, 7 より全ての再現率レベルで精度が改善している. この結果より単語間の関連性にもとづき文書間の類似性を求める本手法が有効であることがわかる. また, ベクトル空間モデルより素朴な EMD がよりよい結果となっている. これは, EMD が輸送問題として距離を計算していることに起因する.

例えば, 検索課題を $(\sqrt{3}/2, 1/2)$ としたとき, ベクトル空間モデルでは $(1, 0)$ と $(1/2, \sqrt{3}/2)$ では同じ類似度の 0.866 となるが, EMD では 0.134 と 0.268 と異なり, $(1, 0)$ の方が検索要求に近くなる. これは $(1, 0)$ では一番目の方が大きい, $(1/2, \sqrt{3}/2)$ では反転していることが理由で

表 11 MAP (表記のみ活用する他の手法との比較)

手法	MAP(LEVEL1)	MAP(LEVEL2)
VSM	0.4334	0.564
Okapi BM25	0.4422	0.5705
EMD	0.4774	0.6092
提案手法	0.5214	0.6244

図 6 LEVEL1 での再現率—精度グラフ（他手法との比較）

図 7 LEVEL2 での再現率—精度グラフ（他手法との比較）

ある。本実験ではこの効果が良い方に働いていると考えられる。

次に、比較手法と提案手法が統計的に有意な差があるか検定を行った。検定方法は参考文献(岸田 2001)を参考にした。以下にその検定方法を述べる。

比較する 2 つの手法を a と b 、検索課題数を K 、検索課題 h の平均精度を v_h とし、MAP を $\bar{v} = K^{-1} \sum_{h=1}^K v_h$ とする。同一の検索課題に対する 2 つの平均精度を比較することになるので、対標本と捉えることができ、検索課題ごとの手法間の差自体を標本と考える。すると、帰無仮説「2 つの手法の平均精度の母平均の差は 0 である」の下に正規母集団を仮定すれば、

$$t = \frac{\bar{v}_a - \bar{v}_b}{\sqrt{\frac{s_a^2 + s_b^2 - 2Cov_{ab}}{K}}} \quad (27)$$

が自由度 $K - 1$ の t 分布に従うことになる. この時, s_a^2 と s_b^2 は平均精度の標本分散であり, Cov_{ab} は v_{ha} と v_{hb} との共分散で, $Cov_{ab} = (K - 1)^{-1} \sum_{h=1}^K (v_{ha} - \bar{v}_a)(v_{hb} - \bar{v}_b)$ となる.

この検定方法を利用し, 今回の実験での LEVEL1 における提案手法と比較手法の中で一番評価が高かった素朴な EMD との評価に有意な差があるかどうか片側検定を行う. a を提案手法, b を素朴な EMD とすると, $s_a^2 = 0.071701$, $s_b^2 = 0.067863$, $\bar{v}_a - \bar{v}_b = 0.044005$, $Cov_{ab} = 0.056419$ となるので, $t = 1.723548$ となる. 自由度 40 の t 分布に従うので, $1.723548 > 1.684$ より, 帰無仮説は棄却され, 5%水準で提案手法と素朴な EMD との差は有意であることがわかる. また, 同様の検定をベクトル空間モデルで行うと $t = 3.01916$, OkapiBM25 では $t = 3.563336$ となり 1%水準で有意な差があることを確認できた. 提案手法が表記に頼る比較手法に比べ統計的に有意な差があることがわかった.

9 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 8.1 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. 平均精度の平均 (MAP) を表 12 に示す. 再現率—精度グラフを図 8, 9 に示す. 表 12 から分かるように, LEVEL1 において 1.58%の精度向上を達成し, LEVEL2 において 3.04%向上している. LEVEL1 においては全ての再現率レベルで提案手法が EMD+CB を上回っているが, LEVEL2 においては再現率レベル 0.3 の点で 0.0001 だけ EMD + CB を下回っている. より詳細に検証するために, 検索課題ごとの平均精度の差をプロットしたものを図 10 に示す. 横軸は本実験における検索課題の番号 (1~41) で, 縦軸は検索課題毎のグラフが上に伸びているほど自動拡張手法によって精度が向上したことを示し, 逆に下にのびているほど精度は低下していることを示している. この結果から検索課題 14 と検索課題 29 の時に精度の低下が著しいことがわかる.

課題 14 は「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」という課題で, 課題 29 は「シフォンケーキの作り方が書かれている文書を探したい.」といった課題である. これらに形態素解析を施し, 検索課題 idf による不要語の削除を行うと, 検索課題 14 は「宮部／

表 12 MAP (自動拡張手法使用と未使用での比較)

手法	MAP(LEVEL1)	MAP(LEVEL2)
EMD + CB	0.5133	0.606
提案手法	0.5214	0.6244

図 8 LEVEL1 での再現率—精度グラフ (自動拡張手法使用と未使用での比較)

図 9 LEVEL2 での再現率—精度グラフ (自動拡張手法使用と未使用での比較)

みゆき／執筆／小説／書評／レビュー」となり，検索課題 29 は「シフォンケーキ／作り方」となる．このうち未定義語は「宮部／シフォンケーキ」であった．

検索課題 14 と検索課題 29 における上位 25 件の順位の変動を図 11 に示す．図 11 において色が付いてる文書が LEVEL1 における正解文書である．一番精度低下が著しかった課題 29 で，EMD+CB において 47 位であった不適合文書が提案手法では 3 位に，5 位であった不適合文書が 4 位になり，適合文書が 3 位から 5 位に下がった．提案手法において 5 位であった高適合文書と 3 位であった不適合文書を図 12, 13 に示す．3 位の文書はケーキ屋の紹介が書かれている文書で，4 位の文書もケーキに関する文書であった．5 位の高適合文書はシフォンケーキの作り方の手順が書いてあった．提案手法では，シフォンケーキがケーキに関する様々な語と高関連であると判定し，ケーキに関する文書が検索の上位にきてしまった．ユーザは「ケーキの作り

図 10 課題別の平均精度の差

課題14		課題29	
自動拡張手法 未使用	自動拡張手法 使用	自動拡張手法 未使用	自動拡張手法 使用
A	NW015004815	A	NW011761975
C	NW006604666	C	NW008528883
A	NW015004807	A	NW011106020
C	NW00087032	C	NW008528885
C	NW001749679	C	NW008528845
A	NW015004806	C	NW008528845
C	NW006604178	C	NW008528846
C	NW006604179	C	NW001542391
C	NW006604183	C	NW008528885
C	NW006604173	C	NW008528895
C	NW006604147	C	NW008528843
A	NW00587254	C	NW007799707
C	NW002782136	C	NW004295761
C	NW006604170	C	NW008528851
C	NW006604294	C	NW008528882
C	NW006604221	C	NW007799656
C	NW006604220	C	NW013569363
C	NW008704194	C	NW007799704
C	NW006604220	C	NW007725141
C	NW001938374	C	NW008528863
C	NW000459982	C	NW007799706
C	NW006603480	C	NW005344045
C	NW008704194	C	NW005039129
C	NW006603419	C	NW004508569
C	NW006604222	C	NW008528864
C	NW008081367	C	NW008528891
C	NW008081367	C	NW007413264
C	NW006604131	C	NW006328470

図 11 課題 14 と課題 29 における上位 25 件のランク付け

方」と入力せず「シフォンケーキの作り方」と入力したのはケーキ全般の情報よりシフォンケーキに絞った情報が欲しいといったことを暗に示していると考えられ、シフォンケーキを定義してしまったことが悪影響を及ぼしてしまった。

シフォンケーキのようなケーキといった一般的な語に比べ、具体的な語はユーザが検索結果を絞るために用いる語であり、具体的な語に対しては、表記に頼る方が有効に働くと考えられる。具体的な語を判別し、なんらかの対応を行うことでこの問題は解決できると考えられる。

1 2. オレンジシフォンケーキ
 1 2. オレンジシフォンケーキ (lunlun さんの紹介です)
 オレンジ (ネーブルか清見オレンジ) の香りを生かした甘さ超控えめのフワフワのシフォンケーキを考えてみました。オレンジクリームを添えて食べたらなおおいしい! 一度上手く焼けたら、やみつきになる美味しさ! 材料薄力粉 1 2 0 グラム卵黄 4 個卵白 6 個グラニュー糖 8 0 グラムサラダオイル 6 5 グラムオレンジの絞り汁 1 0 0 c c オレンジの皮のすりおろし 1. 5 個作り方
 ①薄力粉はふるっておく。
 ②卵白を泡立て器で泡立て、グラニュー糖 4 0 グラムを 2 ~ 3 回に分けて加え、しっかりしたメレンゲを作る。
 ③卵黄にグラニュー糖 4 0 グラムを加え、泡立て器でもったりするまでかき混ぜ、サラダオイル、オレンジの皮のすりおろしと絞り汁の順に入れ、よく混ぜる。
 ④③に①を加え軽く混ぜる。
 ⑤④にメレンゲの 1/3 を入れてむらなく混ぜてから、残り 2/3 を入れ、泡をつぶさないように切るような感じで混ぜ合わせる。
 ⑥シフォンケーキ型に流し入れ、約 1 7 0 ℃のオーブンで 4 5 分焼く。
 ポイントは、卵白の十分な泡立てです。甘さ控えめだから、オレンジクリームを添えてどうぞ!!
 夏みかんをミキサーでジュースにしたものを使えば、夏みかんシフォン! 甘さは、調節を。

図 12 高適合文書

ブティックタイユバン・ロブション
 ブティックタイユバン・ロブションでは、ロブション独自の天然酵母でじっくり焼き上げたパンをはじめ自家製のケーキ、焼き菓子など多彩にそろえております。
 特別な方への贈り物として、ロブションのレシピと吟味した材料で作るフランス伝統の焼き菓子、そしてお茶やジャムなど。本物のフランスを堪能いただける、シャトーレストランのみがご提供する逸品です。
 中略
 タルトレット・オ・ブルダルー
 さっぱりした洋梨のタルト
 モンブラン (季節商品)
 フランス産の栗を使ったクリーム
 中略
 ロブションのレシピと吟味した材料で作るフランス伝統の焼き菓子、そしてお茶やジャムなど本物のフランスをご堪能いただける、シャトーレストランのみがご提供する逸品です。またシャトーレストランのメインダイニングで使用しているオリジナル陶器が、ご家庭でもご愛用いただけます。ロブションの料理にふさわしい質と美しさを追求し、名窯リモージュで焼きあげた食器は、高貴な貴族趣味のなかにも現代感覚を漂わせ、格調高くエレガント。料理の魅力をより一層引き立て、食卓を優雅に演出します。大切な方への贈り物に最適です。
 場所: 恵比寿ガーデンプレイス地下 1 階営業時間: 10:00a.m. ~ 8:00p.m. 電話番号: 03-5424-1345 定休日: 無休

図 13 不適合文書

10 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。さらに、統計的検定を用いることで、提案手法と比較手法の性能の差に有意性があることを確認した。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。

今後の研究課題としては、検索要求内に具体的な語が含まれている場合の対応である。ユーザは検索結果を絞るために具体的な語を用いていると考えられ、これらの語をシソーラスなどをもとに判別し対応を行えば更なる精度向上が期待できる。また、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。

参考文献

- 笠原要, 松澤和光, 石川勉 (1997). “国語辞書を利用した日常語の類似性判別.” 情報処理学会論文誌, **38** (7), pp. 1272–1283.
- 岸田和明 (2001). “検索実験における評価指標としての Mean Average Precision の性質.” 情報処理学会研究報告, **2001** (74), pp. 97–104.
- 小島一秀, 渡部広一, 河岡司 (2002). “連想システムのための概念ベース構成法—属性信頼度の考え方に基づく属性重みの決定.” 自然言語処理, **9** (5), pp. 93–110.
- Miller, G. (1995). “WordNet: A lexical database for English.” *Commun. ACM*, **38** (11), pp. 39–41.
- 奥村紀之, 土屋誠司, 渡部広一, 河岡司 (2007). “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, **14** (5), pp. 41–64.
- Robertson, S., Walker, S., Jones, S., Beaulieu, M., and Gatford, M. (1995). “Okapi at TREC-3.” In *In proceeding of the 3rd Text Retrieval Conference*, pp. 109–126.
- Rubner, Y., Tomasi, C., and Guibas, L. (2000). “The earth mover’s distance as a metric for image retrieval.” *Int. J. Comput. Vision*, **40**, pp. 99–121.
- Salton, G. and Buckley, C. (1988). “Term-weighting approaches in automatic text retrieval.” *Information Processing and Management*, **41** (4), pp. 513–523.
- Salton, G., Wong, A., and Yang, C. (1975). “A Vector space model for automatic indexing.” *Communications of the ACM*, **18** (3), pp. 613–6201.
- 辻泰希, 渡部広一, 河岡司 (2004). “www を用いた概念ベースにない新概念およびその属性獲

得手法.” 第 18 回人工知能学会全国大会論文集, 2D1-02.

Wan, X. and Peng, Y. (2006). “The Earth Mover's Distance as a Semantic Measure for Document Similarity.” In *Proceeding of the 14th ACM international conference on Information and knowledge management*, pp. 301–302.

渡部広一, 奥村紀之, 河岡司 (2006). “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, **13** (1), pp. 53–74.

柳本豪一, 大松繁 (2007). “Earth Mover's Distance を類似度として用いた情報検索.” 平成 19 年度 電気学会全国大会, 3-065.

略歴

藤江 悠五：2007 年同志社大学工学部知識工学科卒業. 2009 年同大学院工学研究科知識工学専攻博士前期課程修了. 同年, パナソニック株式会社入社. 知識情報処理の研究に従事.

渡部 広一：1983 年北海道大学工学部精密工学科卒業. 1985 年同大学院工学研究科情報工学専攻修士課程修了. 1987 年同精密工学専攻博士後期課程中途退学. 同年, 京都大学工学部助手. 1994 年同志社大学工学部専任講師. 1998 年同助教授. 工学博士. 現在同教授. 主に, 進化的計算法, コンピュータビジョン, 概念処理などの研究に従事. 言語処理学会, 人工知能学会, 情報処理学会, 電子情報通信学会, システム制御情報学会, 精密工学会各会員.

河岡 司：1966 年大阪大学工学部通信工学科卒業. 1968 年同大学院修士課程修了. 同年, 日本電信電話公社入社, 情報通信網研究所知識処理研究部長, NTT コミュニケーション科学研究所所長を経て, 同志社大学工学部教授. 2009 年定年退職. 工学博士. 主にコンピュータネットワーク, 知識情報処理の研究に従事. 言語処理学会, 人工知能学会, 情報処理学会, 電子情報通信学会, IEEE(CS) 各会員.

(2008 年 11 月 12 日 受付)

(2009 年 2 月 10 日 再受付)

(2009 年 2 月 16 日 採録)