G-03

音声入力チャットの話題解析による情報検索効率化と内容可視化の研究
Information retrieval efficiency and contents visualization by topic analysis of voice inputting chat

山中 勇矢†
Yuya Yamanaka

1. はじめに

データマイニングなどの IT 技術の発展に伴い，インターネットや企業に蓄積された膨大なデータから，いかに有益な情報を導き出すことができるかが注目されている。多くの企業では，蓄積された膨大なデータから，業務の効率化・コスト削減・顧客の潜在ニーズの発掘など，企業価値向上への様々な活用を試みている。しかし，既に蓄積して活用されているデータ以外にも，企業に有益な情報をもたらすデータは潜在している。その例として，業務・会議・ミーティングにおける会話や紙媒体で保存されている文書など，「デジタル化されていないデータ（アナログデータ）」がある。このようなアナログデータをデジタルデータに変換する方法として，音声認識技術がある。音声認識技術の発展に伴い，話し言葉を比較的高精度でテキスト化することが可能となった。また，キーボードの打鍵を必要としないそのテキスト入力方法は，タイピングが遅い人にとっての作業効率向上やタイピングによる疲れがないといった恩恵をもたらし，大量の文書をデジタル化する方法としても効果的である。こうした方法により，企業は蓄積するデータ量を増加させ，さらなる企業価値の向上に活用することが可能となる。

蓄積される情報量の増加は，企業に様々な企業価値向上の機会を与える。その一方で，情報量が増加するほど，情報の整理は煩雑になり，情報が氾濫していく。そのような情報の中では，人が必要な情報にたどり着く方法が問題となる。必要な情報が存在しているか，あるいは情報が格納された場所がわからない場合，キーワード検索を使用してファイルやフォルダの名前を検索する方法が一般的である。キーワード検索を使用した場合，検索結果としてキーワードを含むファイルやフォルダの名前が表示される。しかし，表示されたファイルの中身が本当に必要な情報を記載しているかどうかは，結局その中身を読んでみるしかないということがしばしばある。また，ファイルの中身が必要な情報を記載していても，ファイルやフォルダの名前が入力したキーワードを含んでいるとは限らない。結果として検索したファイルの中身を読むことに時間を費やしたにもかかわらず，必要な情報にたどり着けなかったという非効率的なことが起こり得る。このような情報検索の非効率性は，情報にたどり着くために費やす時間を延ばしたり，前例のある作業を新規の作業として一から実施するという非効率的な状態を生み出したりする。

本研究では，音声入力によってアナログデータを文書データ化・蓄積する環境を構築し，企業が保有する情報量の増加を支援する。また，増加によって煩雑・氾濫する情報に対して，検索機能による文書データへのアクセス性向上や，文書データを分析することで文書同士の関連付けや内容の可視化を試み，情報検索の効率化をサポートするシステムを開発する。それによって，情報検索の非効率性を解消することを目指す。

2. 使用する手法

2.1 音声認識技術

音声認識技術は，人の発する声をコンピュータに認識させることで，話し言葉をテキストに変換することを可能とする技術である。本研究では，音声認識を使用して蓄積された文書データを分析することを主軸としているため，システムの開発にあたっては既に開発された音声認識技術を取り入れることとする。音声認識技術の利用方法として，音声入力によって文書データを作成することに加え，会話という複数人が行う形態への対応が必要であることから，Web コンテンツとして連携が容易である API による音声認識技術を使用する。本研究では，Google が提供する Web Speech API を使用してシステムの開発を進める。

2.2 形態素解析

形態素解析（Morphological Analysis）は，自然言語処理の基礎技術の一つであり，自然言語で書かれた文を「言語が意味を持つ最小単位」である形態素に分割し，品詞を見分ける作業である。形態素解析を行うためには様々なツールが提供されているが，今回開発中のシステムでは，無料で配布されている日本語形態素解析ツール MeCab を使用する。MeCab は，高い判別精度に加えて実行速度が速く，形態素を名詞などの品詞・固有名詞などの細分類・単語の活用形・原形・読みまで判別することができる。さらに，新語への対応が実施される辞書 mecab-ipadic-neologd も配布されており，商品名や映画・ドラマ・アニメ・書籍のタイトルなど様々な単語を正確に分かち書きすることが可能である。

2.3 TF-IDF[1]

TF-IDF（Term Frequency - Inverse Document Frequency）とは，「文書内に含まれる単語の特徴の強さを表す数値」である。ある文書中における単語の出現頻度を「Term Frequency」と呼び，文書を構成する単語総数に対して，ある単語が占める割合を表す。また，文書のグループ全体を通した単語の出現頻度は「Document Frequency」と呼び，文書総数に対して，ある単語を含む文書総数が占める割合となり，「Inverse Document Frequency」は「Document Frequency」の逆数である。TF-IDF による数値の算出は次のような考え方となる。

TF-IDF の値 = (文書における単語 A の出現頻度) * log(文書総数 / 単語 A が出現した文書数)

形態素解析を行った単語に TF-IDF を適用することにより，文書における特徴語を抽出することができる。

2.4 N-gram[2]

N-gram とは，テキストで隣り合った N 文字のことを指しており，文書の類似度を調べるために N-gram を利用す

†京都情報大学院大学，The Kyoto College of Graduate Studies for Informatics

ることが可能である。異なる 2 つの文書の N-gram を総当たりで比較すれば，文節の順序が異なっていても，出現する単語の種類や頻度を比較することができるという考え方で，例えば，「今日は快晴です。」という文書に対して，N-gram を 3 文字（3-gram）とすると，「今日は」「日は快」「は快晴」「快晴で」「晴です」「です。」に分解される。このように分解した文書と同様に区切った別の文書を比較し，どの程度一致しているかを算出することで，文書データ同士の類似度を算出することができる。

3. 分析の方法

3.1 分析の進め方について

蓄積した文書データ同士の関連付けと内容の可視化から，情報検索の効率化に結び付けるために，MeCab による形態素解析を通して，文書データを名詞・動詞・形容詞などの品詞ごとに分類する。文書データを構成する名詞・動詞・形容詞の単語を対象に，重要度をスコア形式で評価する。例えば，文書データ A の名詞単語「蓄積」がスコア 10，「形態素解析」がスコア 20 ならば，文書データ A において「形態素解析」の方が重要度は高いという判定になる。スコアリングの方法は，TF-IDF を参考にして行うが，後述する TF-IDF における問題点から，IDF に代わる補正値を新たに算出する。加えて，単語の表記や文字数に着目した補正値の算出を行い，名詞・動詞・形容詞の品詞ごとに評価対象となる単語についてスコアリングを行う。作成した名詞のスコアデータから，特徴語として文章を強く表す語句（タグ）を抽出し，N-gram・タグの一致・単語の一致から文書データ同士の類似度を評価する。

類似度の評価を基にした文書データ同士の関連付けは，「過去に蓄積した似た文書データ」を抽出する役割を持つ。過去の話題から参考になりそうな話題を自動的に抽出し表示することで，検索を行うことなく似た内容の文書データの参照を可能にする。これを実現するためには，正確に「似た内容の文書データは類似度を高く，異なった内容の文書データは類似度を低く」評価する必要がある。そのために，スコアリングの評価方法を基に話題解析の機能を実装し，似た内容の文書データと異なった内容の文書データを蓄積して類似度の評価を分析する。

内容の可視化の分析では，スコアリングを基にした名詞・動詞・形容詞についてのスコアリング表をグラフと組み合わせて表示し，文書データの内容に目を通すことなく内容を推察可能であるか試みる。

類似度の高い文書データの抽出と内容の可視化による推察を組み合わせることによって，情報へのアクセス性の向上と情報が必要か否かの判断の高速化を図り，情報検索の効率化に結び付ける。

3.2 TF の算出方法

TF の値の算出は，文書における単語 A の出現頻度（単語 A の出現回数÷単語の出現総数）で行う。この際，TF は名詞・動詞・形容詞でそれぞれ分けて算出する。TF の算出は，新聞記事のようにある程度の長い文書なら重要な単語が繰り返し出現するという TF の仮定が有効である。[3]しかし，文書データが短文や人によって特徴の出やすい会話形式である場合，単語によって TF の結果が偏ってしまう。そのため，より高精度な特徴語抽出を行うためには，IDF による補正が重要となる。

3.3 IDF の算出方法

IDF は，文書総数に対して，単語 A を含む文書数が占める割合の逆数（文書総数÷単語 A が出現した文書数）である。開発中のシステムでは，文書総数は流動的に増減するため，同一の文書データであっても，どのタイミングで蓄積されるかにより IDF の値が異なるという欠点がある。例えば，文書データ A が「音声入力チャットの話題解析」という内容であった場合，MeCab による形態素解析を行うと「音声｜入力｜チャット｜の｜話題｜解析」と分かち書きされる。文書データ A から文書データ E のうち，「音声」を含んだ文書データが文書データ A・文書データ B・文書データ E の 3 つだったとする。この場合，文書 A の蓄積されるタイミングおよび文書総数と「音声」の出現数をもとにした名詞での TF-IDF の数値は次のようになる。

表 1　TF-IDF の算出

TF	文書総数	蓄積文書	「音声」を含む文書数	IDF	TF-IDF
0.2	1	A	1	0	0
	2	B・A	2	0	0
	3	B・C・A	2	0.4055	0.0811
	4	B・C・D・A	2	0.6931	0.1386
	5	B・C・D・E・A	3	0.5108	0.1022

表 1 から，文書が蓄積されるタイミングによって TF-IDF の数値が変化することがわかる。これによって，4 番目に登録された場合に「音声」は文書データの特徴語として抽出されるが，1 番目や 2 番目に登録された場合は特徴語として抽出されないというケースが発生する。開発中のシステムでは，文書間の類似度を算出するためにタグ（特徴語）を比較する必要があるため，文書データは定まった評価方法で分析される必要がある。人が文書データ A と文書データ B を見て，両方とも「音声」が特徴語であると感じても，IDF によって 1 番目に蓄積された文書データ A では特徴語となり，その後に登録される文書データ B では特徴語として抽出されず，評価の妥当性が欠ける恐れがある。評価に妥当性を持たせるため，IDF に代わって独自に作成した補正値の算出方法を次節で述べる。

3.4 IDF に代わる補正値の算出方法

開発中のシステムにおける IDF の欠点は，数値が流動的に変化することにある。これを解決するために，名詞について「一般的によく使われる単語（基本単語）」を導き出し，それを基にして IDF に代わる補正値を作成する。基本単語の作成方法として，無料で配布されている livedoor のニュースコーパスと Wiktionary の日本語の基本語彙 1000 をサンプルに使用して作成する。日本語の基本語彙 1000 のうち，「今日」「子供」「健康」などの名詞 545 語を抽出し，この 545 語を基本単語のベースとする。次に，ニュースコーパスから出現頻度の高い名詞を抽出する。ニュースコーパスはトピックニュース・Sports Watch・IT ライフハック・家電チャンネル・MOVIE ENTER・独女通信・エスマックス・livedoor HOMME・Peachy の全 9 カテゴリ，7,366 個のテキストファイルで構成されている。全テキストファイルの名詞 80,658 語の中から 1 回しか出現しなかった名詞 38,125 語を除外し，2 回以上出現した名詞 42,533 語のうち出現頻度の高かった上位 1%の 425（同率含む 428）

語を抽出する。

表 2　サンプルからの名詞抽出方法

項目	内容	数値
1	日本語の基本語彙 1000 の名詞数	545 語
2	ニュースコーパスのカテゴリ数	9 カテゴリ
3	ニュースコーパスのテキストファイル数	7,366 個
4	3 内に存在する名詞数	80,658 語
5	4 のうち 1 度しか出現しなかった名詞数	38,125 語
6	4 のうち 2 回以上出現した名詞数	42,533 語
7	6 のうち出現頻度上位 1%（同率含む）	428 語

続いて，ニュースコーパスのカテゴリごとに TF-IDF による特徴語抽出を行う。以下は，例として独女通信のカテゴリ内で特徴語を抽出した結果，TF-IDF の数値が上位 20 位までの表である。

表 3　「独女通信」カテゴリ内での TF-IDF の算出

順位	単語	TF-IDF	順位	単語	TF-IDF
1	結婚	0.0090	11	独女	0.0040
2	自転車	0.0069	12	子供	0.0039
3	女	0.0059	13	好き	0.0039
4	男性	0.0059	14	夫	0.0039
5	男	0.0056	15	女子	0.0038
6	恋愛	0.0052	16	映画	0.0037
7	相手	0.0047	17	友達	0.0034
8	仕事	0.0046	18	人	0.0034
9	女性	0.0045	19	話	0.0034
10	メール	0.0043	20	母	0.0033

他 8 カテゴリについても同様の表を作成し，表 2 で作成した 428 語がカテゴリごとの特徴語 20 語と日本語の基本語彙 1000 に含まれているかマッピングを行い，出現回数を求め，以下の算出式に従って補正値を決定する。

補正値 = log{1 / (文書全体における単語の出現頻度 * 単語のカテゴリ出現頻度)}

表 4　基本単語の補正値の算出例

単語	文書全体における単語の出現頻度①	単語のカテゴリ出現頻度②	①×②③	③の逆数④	log（④）補正値
人	0.3808	0.7	0.2665	3.7514	1.3221
日本	0.2389	0.4	0.0955	10.4630	2.3478
機能	0.1768	0.4	0.0707	14.1327	2.6484
写真	0.1307	0.4	0.0522	19.1225	2.9508
情報	0.2367	0.2	0.0473	21.1180	3.0501
仕事	0.1490	0.3	0.0447	22.3619	3.1073

表 4 で作成した基本単語の補正値は，1.3221…から 5.5689…の間の値をとり，どのカテゴリにも存在しなかった日本語の基本語彙 1000 の名詞に関しては補正値として 6 を設定する。

動詞については，日本語の基本語彙 1000 のうち，「始める」「もらう」「帰る」などの動詞 175 語を抽出し，この 175 語を基本単語のベースとする。次に，ニュースコーパスから出現頻度の高い動詞を抽出する。全テキストファイルの動詞 4,806 語の中から 1 回しか出現しなかった動詞 1,164 語を除外し，2 回以上出現した動詞 3,642 語のうち出現頻度の高かった上位 3%の 109 語を抽出する。以降は，名詞と同様の手順で基本単語の補正値を算出する。補正値は，0.5095…から 5.3812…の間の値をとり，どのカテゴリにも存在しなかった日本語の基本語彙 1000 の動詞に関しては補正値として 6 を設定する。ただし，「する」「できる」「いう」は日本語の基本語彙 1000 には存在しないが出現頻度が極めて高いことから，これらの単語が特徴語とならないよう極めて低い補正値（する：0.01，できる：0.05，いう：0.05）を設定する。

形容詞については，日本語の基本語彙 1000 のうち，「嬉しい」「広い」「白い」などの形容詞 67 語を抽出し，この 67 語を基本単語のベースとする。次に，ニュースコーパスから出現頻度の高い形容詞を抽出する。全テキストファイルの形容詞 722 語の中から 1 回しか出現しなかった形容詞 171 語を除外し，2 回以上出現した形容詞 551 語のうち出現頻度の高かった上位 10%の 55 語を抽出する。以降は，名詞・動詞と同様の手順で基本単語の補正値を算出する。補正値は，1.1214…から 6.2374…の間の値をとり，どのカテゴリにも存在しなかった日本語の基本語彙 1000 の形容詞に関しては補正値として 7 を設定する。ただし，「ない」「いい」「やすい」は日本語の基本語彙 1000 には存在しないが出現頻度が高いことから，これらの単語が特徴語とならないよう低い補正値（ない：0.05，いい：0.05，やすい：0.1）を設定する。

3.5　英字または日本語による補正値の算出方法

MeCab による形態素解析を行う場合に，「Web（ウェブ）」や「IT（アイティー）」などの英字で構成された単語には注意が必要である。例えば，「Web」は「web」「WEB」「Web」という表記が，「IT」は「it」「IT」「It」という表記が考えられる。それぞれについて形態素解析を行うと，結果は次のようになる。

表 5　大文字・小文字による形態素解析のブレ

単語	品詞	品詞細分類 1	品詞細分類 2
web	名詞	固有名詞	人名
WEB	名詞	固有名詞	組織
Web	名詞	固有名詞	一般
it	名詞	固有名詞	人名
IT	名詞	固有名詞	一般
It	名詞	固有名詞	人名

表 5 のように，英字の大文字と小文字の違いによって，固有名詞のうち「一般」「人名」「組織」であるかどうかが異なる。開発中のシステムで使用している Web Speech API では，英字で構成された単語はすべて小文字でテキスト化されてしまうため，正確に分析させるためには手作業での修正が必要となる。このような理由から，日本語の固有名詞と比較し，英字は表記による形態素解析のブレが大きいため，日本語の単語よりも補正値を低く設定する。ただし，この補正値は補助的な数値であり，大きく差をつけてしまうと特徴語の抽出に偏りが出るため，「英字単語と日本語単語が同一の評価を得た場合に日本語単語を優先する」程度に留めておくこととする。そのため，英字単語の補正値を 1 として，日本語単語の補正値を 1.1 に設定する。

3.6　文字数による補正値の算出方法

文書データの特徴語として「心」「精神」「精神科」「精神病院」「スピリチュアルカウンセラー」の 5 語を並べた場合，「心」のように解釈が複数ある単語は，精神を表しているのか，感情を表しているのかなど，受け手によ

って印象が大きく異なることがある。また，「精神」においても同様で，忍耐などを連想するのか，精神科や精神病院のような特定の施設を連想するのかが受け手によって異なる。これらの単語と比較すると，「精神」＋「病院」という単語同士が連結して一語を構成する「精神病院」や「スピリチュアルカウンセラー」などは意味の特定が容易である。「心」や「精神」が特徴語となるより，「精神病院」や「スピリチュアルカウンセラー」が特徴語として抽出された方が相応しいことから，文字数が多い単語ほど受け手がその内容を特定しやすいと仮定し，次のように文字数での補正値を算出する。

表 6　文字数による補正値の算出方法

文字数	補正値 = log(文字数)
1	0
2	0.6931
3	1.0986
4	1.3863
5 以上	1.6094

文字数は，半角文字・全角文字問わず 1 文字で文字数 1 とする。文字列での補正値は「スピリチュアルカウンセラー」「プレゼンテーション」「ビジネスモデル」などのような，カタカナ語の補正値が高くなることが多い。そのため，補正値による極端な差が発生しないよう 5 文字以上の単語は同一の補正値を設定する。

3.7　名詞のスコアリング

名詞のうち，「a」や「A」など英字 1 字で構成された単語や，「主」「心」「我」など日本語 1 字で構成された単語では文書データの内容を把握することが困難である。そのため，英字・日本語問わず 1 字で構成された名詞はスコアリングから除外する。

MeCab の品詞体系のうち，品詞細分類 1 が「ナイ形容動詞語幹」「接続詞的」「数」「引用文字列」「接尾」「非自立」「代名詞」「副詞可能」「特殊」に該当する名詞は，それ単体では文書データの内容を把握することが困難であるためスコアリングから除外する。品詞細分類 1 が除外対象の単語例は次の通りである。

表 7　除外する品詞細分類 1 に該当する単語例

品詞細分類 1	単語例
ナイ形容動詞語幹	問題
接続詞的	対・兼
数	1・2・3
引用文字列	いわく
接尾	殿・部・的・課
非自立	ため・ところ・こと
代名詞	私・彼・彼女
副詞可能	日々・今・現在
特殊	（大き）さ

MeCab での形態素解析では，数字を含んだ単語が固有名詞として判別されるという欠点がある。例えば，「平成 29 年」などの和暦，「2017 年」などの西暦，「8 月 20 日」などの日付，「-10」「1.5cm」「1000 円」「10%」などの記号や単位を含んだものがある。数字を含んで構成された固有名詞は，正規表現を用いることでスコアリングから除外する。ただし，「Windows10」や「HTML5」のような 2 文字以上の英字＋数字で構成された固有名詞は除外しない。

MeCab は，英字の大文字・小文字による判別のブレの他に，ひらがな表記の単語を漢字単語として判別することがある。例えば，「（そんなことは）ありません」や「（雑談）しながら」などの単語は，「ありません（有馬線）」や「しながら（品柄）」として判別される。文書データに出現した文字列（表層形）がひらがなであるが，形態素解析の結果で表示された単語の原形が漢字である名詞についてはスコアリングから除外する。

スコアリングから除外するにあたって，規則に当てはまらないものは別途「除外リスト」を作成して対応する。開発中のシステムでは「自分」「自分自身」「自身」のみを設定している。

名詞のスコアリングでは，単語の TF に補正値を考慮して算出する。補正値は，MeCab の品詞体系のうち品詞細分類 1 によって判断し，補正値の値は「固有名詞＞一般＞サ変接続＞形容動詞語幹＞（その他）＞3 章 4 節で作成した基本単語の補正値」の順に大きくする。また，固有名詞については，「一般」「組織（日本語）」「組織（英字）」「地名」「人名（日本語）」「人名（英字）」で値に差をつける。各補正値は次の通りである。

表 8　名詞の補正値一覧

品詞細分類 1	補正値
固有名詞	-
（固有名詞）一般	20
組織（日本語）	15
組織（英字）	6
地域	15
人名（日本語）	10
人名（英字）	6
一般	10
サ変接続	9
形容動詞語幹	8
（その他）	7
基本単語	1.3221 ～ 6

表 8 の補正値と，3 章 2 節から 3 章 6 節までで述べた算出方法を組み合わせた名詞のスコア算出式は次の通りである。

スコア = TF * 補正値
TF = 単語の出現回数 / 全ての単語の総出現回数
補正値 = (単語の補正値 * 10) * 英字・日本語の補正値 * 文字数の補正値
単語の補正値：固有名詞・一般・サ変動詞・形容動詞語幹・基本単語による補正値
英字・日本語の補正値：英字 1・日本語 1.1
文字数の補正：log(文字数)

上記の算出式を使用して，評価対象の名詞すべてに対してスコアリングを行う。

名詞ごとにスコア算出式を用いて評価したスコアを比較して，上位 5 語をその文書データのタグ（特徴語）として設定する。

3.8　動詞のスコアリング

MeCab の品詞体系のうち，品詞細分類 1 が「非自立」「接尾」に該当する動詞は，それ単体では文書データの内容を把握することが困難であるためスコアリングから除外する。品詞細分類 1 が除外対象の単語例は次の通りである。

表 9　除外する品詞細分類 1 に該当する単語例

品詞細分類 1	単語例
非自立	れる・いる
接尾	れる・られる

名詞と同様に，単語の TF に補正値を考慮して算出する。スコアリングの際には，動詞は文書中に出現した単語（表層形）から MeCab による形態素解析を通して原形に変換する。例えば，「○○を受けて」の「受け」は「受ける」に変換する。動詞においては英字・日本語の補正値と文字数による補正値は含めない。以上を踏まえた動詞のスコア算出式は次の通りである。

スコア = TF * 補正値
TF = 単語の出現回数 / 全ての単語の総出現回数
補正値 = 単語の補正値 * 10
単語の補正値：基本単語による補正値・基本単語以外の補正値（一律 10）

上記の算出式を使用して，評価対象の動詞すべてに対してスコアリングを行う。

3.9　形容詞のスコアリング

MeCab の品詞体系のうち，品詞細分類 1 が「非自立」「接尾」に該当する形容詞は，それ単体では文書データの内容を把握することが困難であるためスコアリングから除外する。品詞細分類 1 が除外対象の単語例は次の通りである。

表 10　除外する品詞細分類 1 に該当する単語例

品詞細分類 1	単語例
非自立	づらい・がたい・よい
接尾	ったらしい・っぽい

名詞・動詞と同様に，単語の TF に補正値を考慮して算出する。スコアリングの際には，形容詞は文書中に出現した単語（表層形）から MeCab による形態素解析を通して原形に変換する。例えば，「多く」は「多い」に変換する。形容詞においては英字・日本語の補正値と文字数による補正値は含めない。以上を踏まえた形容詞のスコア算出式は動詞と同様である。上記の算出式を使用して，評価対象の形容詞すべてに対してスコアリングを行う。

3.10　N-gram（文字長）での類似度の評価方法

ある文書データを取り出し，N-gram を用いて既に蓄積した文書データとの類似度を評価する。N-gram では，N の値が大きいほどモデルの精度が落ちる。また，N に 1 を用いた unigram（ユニグラム）では，比較対象が 1 文字のため類似度が高くなりやすいという特徴がある。ここでの N の値は 3 を用いた trigram（トリグラム）で類似度の評価を行う。

3.11　タグ（特徴語）での類似度の評価方法

文書データに付与したタグが，他の文書データのタグとどの程度一致しているかによって類似度を評価する。タグにはスコアとタグ 5 語のスコア合計に占める割合（タグ A のスコア÷タグの総スコア）を保持させ，比較対象の文書データのタグと一致した場合に保持した割合を加算する。

表 11　文書データ A と文書データ B のタグ比較

文書データ A			文書データ B
タグ	スコア	割合	タグ
グループワーク	2.83	0.17	グループ
システム	4.27	0.25	システム
ビジネス	3.05	0.18	ビジネスモデル
学習	3.29	0.19	期末試験
講義	3.57	0.21	講義

表 11 では，文書データ A と文書データ B で一致するタグが「システム」と「講義」である。この場合の類似度は次のように算出する。

類似度 = (システムの割合 + 講義の割合) * 100
= (0.25 + 0.21) * 100
= 46%

この場合，文書データ A から比較した文書データ B との類似度は 46%と評価される。これと同様の計算を他の文書データとも行い，類似度が高い文書データを抽出する。タグの類似度を算出することで，文書を特徴づける単語がどの程度類似しているかがわかる。

3.12　構成要素（評価対象の名詞）での類似度の評価方法

類似度算出の方法はタグによる評価と同様であるが，こちらは文書データ内に含まれる構成要素（評価対象の名詞）で類似度を算出する。例えば，文書データ A 内にスコアリングの対象の名詞が 100 語存在し，文書データ B 内にスコアリングの対象の名詞が 120 語存在した場合，文書データ A の 100 語と文書データ B の 120 語をマッチングさせ，一致した単語の保持する割合（単語 A のスコア÷文書データ A 内の単語スコアの総合計）を加算していく。

表 12　文書データ A と文書データ B の名詞比較

文書データ A			文書データ B
名詞	スコア	割合	名詞
CSS	0.26	0.0024	グループ
HTML	1.11	0.0099	システム
グループワーク	2.83	0.0252	ビジネスモデル
コミュニケーション	2.12	0.0189	期末試験
システム	4.27	0.0381	講義
企業	1.52	0.0136	大学院
学習	3.29	0.0294	学習
実習	1.10	0.0098	情報処理
情報処理試験	1.41	0.0126	HTML
演習	1.10	0.0098	CSS

表 12 では，文書データ A と文書データ B で一致する名詞が「CSS」「HTML」「システム」「学習」である。この場合の類似度は次のように算出する。

類似度 = (CSS の割合 + HTML の割合 + システムの割合 + 学習の割合) * 100
= (0.0024 + 0.0099 + 0.0381 + 0.0294) * 100
= 7.98%

タグによる評価が文書の特徴の類似度を示すものであるなら，構成要素での評価は文書データの中身がどれだけ類似しているかを示す。

3.13　文書データを構成する単語の表示方法

文書データ内における評価対象の名詞・動詞・形容詞で，

各品詞のスコア上位 30 位までを開発中のシステム画面に表示する。また，表示項目は「順位」「単語」「出現回数」「出現回数のグラフ」「スコア」「スコアのグラフ」とする。これによって文書データ内の内容を読むことなく，内容を推察することができるかを試みる。

4. 分析と結果

4.1 同じ項目における分析と結果

分析を行うためのサンプルとして，2 つの「形態素解析」についての説明文（分析サンプル文章 1[2]・分析サンプル文章 2[4]）を音声入力チャットで文書データ化し，開発中のシステムから分析結果を取得する。ここでのサンプルは，分析サンプル文章 1 は詳細な説明，分析サンプル文章 2 は 1 と比べて端的に説明された文章である。「分析サンプル文章 1」のタグは「形態素解析」「単語」「品詞」「日本語」「かな漢字変換」となった。また，「分析サンプル文章 2」と比較した場合，N-gram での類似度は 30%未満，タグでの類似度は 57.31%，構成要素での類似度は 56.95%という結果になった。

次に，「分析サンプル文章 2」を分析対象として，「分析サンプル文章 1」と比較した場合の分析結果を取得する。「分析サンプル文章 2」のタグは「形態素解析」「言語モデル」「形態素」「かな漢字変換」「テキストマイニング」となった。また，N-gram での類似度は 39.57%，タグでの類似度は 63.5%，構成要素での類似度は 71.85%という結果になった。

表 13　類似度の評価結果

	「分析サンプル文章 1」を「分析サンプル文章 2」と比較	「分析サンプル文章 2」を「分析サンプル文章 1」と比較
N-gram での評価	30%未満	39.57%
タグでの評価	57.31%	63.5%
構成要素での評価	56.95%	71.85%

4.2 異なる項目における分析と結果

サンプルとして，「機械学習」についての説明文（分析サンプル文章 3[2]）を音声入力チャットで文書データ化し，分析結果を取得する。「分析サンプル文章 3」のタグは「機械学習」「ベイジアンフィルタ」「学習」「データ」「迷惑メール」となった。また，「分析サンプル文章 1」および「分析サンプル文章 2」と比較した場合，N-gram での類似度は 30%未満，タグでの類似度は一致するタグが全くない，構成要素での類似度は 30%未満という結果になった。

4.3 話題解析の結果

分析サンプル文章 1～3 の次に，著者が大学院での講義の受講状況や学習の経過報告について記した報告資料「報告書：2015 年 12 月」というタイトルの Word 形式の文書ファイルを，音声入力チャットを使って文書データとして蓄積する。さらに，開発中のシステムには同様に文書データ化した「報告書：2015 年 10 月」「報告書：2015 年 11 月」が蓄積されている。「報告書：2015 年 12 月」を分析対象として話題解析を行った開発中システムの分析結果は図 1 の通りである。

タグは「システム」「知的財産権」「ビジネス」「経営資源」「html」となった。また，類似度での評価は，先に蓄積した「分析サンプル文章 1」「分析サンプル文章 2」「分析サンプル文章 3」はすべての項目で類似度 30%以上とは判断されなかった。N-gram での評価では，「報告書：2015 年 10 月」が 32.19%，「報告書：2015 年 11 月」が 31.88%の類似度であると評価された。タグの評価では，「報告書：2015 年 10 月」が 43.18%の類似度であると評価され，「報告書：2015 年 11 月」は類似度が 30%未満となった。構成要素での評価では，「報告書：2015 年 11 月」が 49.84%，「報告書：2015 年 10 月」が 45.08%の類似度であると評価された。

表 14　「報告書：2015 年 12 月」の分析結果

	「報告書：2015 年 10 月」	「報告書：2015 年 11 月」
N-gram での評価	32.19%	31.88%
タグでの評価	43.18%	30%未満
構成要素の評価	45.08%	49.84%

開発中のシステムによる「報告書：2015 年 12 月」の名詞・動詞・形容詞のスコアリング表示画面は図 2 の通りである。

5. 分析の結果に対する考察

5.1 同じ項目における分析結果に対する考察

「形態素解析」という単語について説明した，異なる 2 つの文献からの引用を基に分析を行った。この分析では，「分析サンプル文章 1」から「分析サンプル文章 2」を比較した場合と，「分析サンプル文章 2」から「分析サンプル文章 1」を比較した場合とで類似度が大きく異なっていることがわかった。

表 13 の結果になったのは，「分析サンプル文章 1」が「分析サンプル文章 2」の内容を包括的に含んでいるからだと考えられる。形態素解析について，「分析サンプル文章 1」は詳しく説明しており，「分析サンプル文章 2」は端的に説明している。これら 2 つの文書データの類似度を算出した場合，端的に説明している「分析サンプル文章 2」の内容は「分析サンプル文章 1」に含まれており，逆に，「分析サンプル文章 1」の内容に対して「分析サンプル文章 2」の内容が不足しているからだと考えられる。そのため，一方から比較すると類似度が低く，もう一方から比較すると類似度が高くなっているということが考察される。これら 2 つの文書の関係は以下のような図と解釈できる。

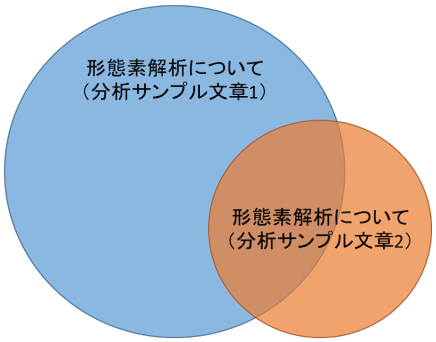

図 3　2 つの文書データにおけるベン図

このことから，内容が短い文書データほど類似度の増減の振れ幅が大きく，極端な数値になりやすい。また，類似度の信頼性に欠けることが考察される。

5.2 異なる項目における分析結果に対する考察

同じ IT 分野で「形態素解析」とは異なる単語「機械学習」についての説明を文書データ化して分析を行った結果，N-gram・タグ・構成要素の全ての評価で類似度 30%未満という数値になった。これによって，異なる，あるいは関係性が薄い文書データについては類似した文書から除外するフィルタが正常に働いていると考えられる。加えて，「報告書：2015 年 10 月」「報告書：2015 年 11 月」「報告書：2015 年 12 月」を含めた分析においても類似文書として抽出されなかったため，スコアリングおよび類似度の評価には一定の成果が得られたと考察される。

5.3 話題解析の結果に対する考察

話題解析に使用した「報告書：2015 年 10 月」「報告書：2015 年 11 月」「報告書：2015 年 12 月」の 3 種の文書データは，著者が大学院での講義の受講状況や学習の経過報告について記している報告資料を文書データとして蓄積したものである。そのため，内容としては同じ趣旨の報告書となる。結果として「報告書：2015 年 12 月」を分析対象とした分析結果は，類似度が「報告書：2015 年 11 月」に対するタグでの評価以外は 30%を超える数値となり，類似度の評価は正しく動作していると考察される。

N-gram での評価は，N-gram の性質から文書のコピーペーストを見分けるために用いられる方法であるため，この類似度が高くなるということは，文書データ同士がそっくりであるということである。

タグでの評価は，文書データを表す特徴語の一致によって算出しているため，同じ内容を記した文書データ同士でも，特定の単語の出現回数によって大きく左右されたり，出現回数が多い＝重要度が高いというわけではなかったり，文書作成者の意図した特徴語とはならないことがあるため，類似度の数値としてはブレが大きいと考えられる。今回使用した「報告書：2015 年 12 月」においても，私がタグを自分で作成するなら「報告書」という単語を使用するが，文書中に「報告書」という単語は存在しないため，システム上タグとして登録されることはない。このようなことから，精度を上げるためには文書タイトルを考慮する必要性も考えられる。

構成要素での評価は，文書データに存在する名詞の一致によって評価するため，文書がそっくりなほど評価が高い N-gram や，ブレが大きいタグでの評価と比べると最も信頼の高い評価方法であると考えられる。これまでの分析からも類似度は，構成要素での評価＞タグでの評価＞N-gram での評価（ただし，タグでの評価のブレは大きい）ということが見て取れる。構成要素での評価が高いということは，それだけ内容が似通っているということを表していると考察される。

名詞・動詞・形容詞のスコアリング表は，人による推察は必要であるものの，文書タイトルである「報告書：2015 年 12 月」と，名詞のスコア上位 30 語「システム」「知的財産権」「ビジネス」「経営資源」「html」「中間テスト」「ビッグデータ」「情報処理試験」「グループワーク」「javascript」「ゼミ」「期末試験」「講義」「分野」「erp」「京都情報大学院大学」などの単語，動詞の「学ぶ」，形容詞の「幅広い」などを組み合わせて解釈すれば，どのような内容の報告書か触り程度は把握が可能ではないだろうか。閲覧者の知識にも大きく影響はされるだろうが，蓄積された文書データに目を通すかどうかの振るい分け程度の効果は期待できるのではないかと考察される。

6. おわりに

分析結果と考察から，文書データの名詞・動詞・形容詞に対してスコアリングを行うことで，類似度の高い文書データを評価して情報検索の効率化が可能であることがわかった。また，開発中のシステムでは蓄積した文書データに対して検索機能を実装し，多角的な検索手段を用いた情報検索の効率化を図ることでさらなる成果が得られるのではないかと考えられる。

一方で，文書データの内容を可視化するという点については，スコアリング表とグラフを組み合わせることで可視化を試みたが，それ単体で内容を把握できるほどの効果は得られなかったと言える。

開発中のシステムはまだ発展途上の段階であり，サービスとして提供するには至っていない。情報検索の効率化や文書データの内容可視化もさることながら，本システムでは誰が会話や文書をデータ化して蓄積したかが明瞭になるため，発言などの責任の所在が非常にシビアになる。そのため，蓄積されたデータに対するアクセス権限やセキュリティ対策について慎重に対策しなければならない。現段階では，これらが万全な状態ではないため，これから開発に注力する必要があると考えられる。

今後は，先に述べたセキュリティ面の実装を含めて，タグの抽出と類似度評価のさらなる精度向上，文書データの可視化を進めていき，業務支援サービスとして提供できる質に達したシステム開発を目指す。

参考文献

[1] 斉藤 常治，高橋 佑幸，PHP による機械学習入門，p.307，株式会社リックテレコム，2014.

[2] クジラ飛行机，いまどきのアルゴリズムを使いこなす PHP プログラミング開発テクニック，p.203，p.292-p.293，ソシム株式会社，2016.

[3] 井前 吾郎，奈須 庄建，淡谷 浩平，重野 寛，松下 温，新聞記事のマイニング結果についての視覚化手法の提案，情報処理学会第 64 回（平成 14 年）全国大会 3C-05，2002.

[4] クジラ飛行机，Python によるスクレイピング＆機械学習 開発テクニック BeautifulSoup、scikit-learn、TensorFlow を使ってみよう，P.256，ソシム株式会社，2017.

[5] 宮崎 将隆，川端 豪，Wikipedia ページへの tfidf 法の適用 A Study of "tfidf" Measure using Wikipedia Statistics，情報処理学会，研究報告音声言語情報処理 Vol.2009-SLP-77 No.2，2009.

[6] 関 洋平，原田 賢一，tf/idf 重み付けに基づく動的文書作成，情報処理学会研究報告デジタルドキュメント 31-4，2001.

[7] 田崎 勝也，コミュニケーション研究のデータ解析，ナカニシヤ出版，p.189-p.200，2015.

[8] 連 燦紅，劉 健全，陳 漢雄，古瀬 一隆，検索語の重要性を考慮した重み付け手法を用いたウェブ検索支援の提案，情報処理学会創立 50 周年記念（第 72 回）全国大会 1R-5，2010.

ルームタグ

タグ1			タグ2			タグ3			タグ4			タグ5		
単語	出現回数	スコア	単語	出現回数	スコア	単語	出現回数	スコア	単語	出現回数	スコア	単語	出現回数	スコア
システム	8	3.48	知的財産権	3	3.03	ビジネス	6	2.61	経営資源	3	2.61	html	3	2.37

類似度の評価

1．N-gramでの評価

ルームID	ルーム名	所属グループ	作成日時	類似度
1	報告書：2015年10月	解析テスト用のグループ	2017-06-30 21:13:46	32.19%
2	報告書：2015年11月	解析テスト用のグループ	2017-06-30 21:14:21	31.88%

2．タグでの評価

ルームID	ルーム名	所属グループ	作成日時	一致したタグ	類似度
1	報告書：2015年10月	解析テスト用のグループ	2017-06-30 21:13:46	システム ビジネス	43.18%

3．構成要素での評価

ルームID	ルーム名	所属グループ	作成日時	一致した単語	類似度
2	報告書：2015年11月	解析テスト用のグループ	2017-06-30 21:14:21	css erp html it javascript sap web ウェブ グループ グループワーク システム セオリー ゼミ データベース ビジネス ビジネスシステム ビジネスモデル プレゼンテーション プログラミング プロジェクト マネジメント メンバー リーダーシップ レポート 中間 京都情報大学院大学 企業 作成 内容 分析 分野 前月 勇矢 受講 報告 大学院 学習 実施 実践 山中 形式 手法 新規 期末試験 期間 来季 来期 業務 概論 構築 機会 準備 演習 状況 理事長 理解 生活 知的財産権 知識 研修 研究 管理 紹介 経営戦略 経過 統合 総合職 総評 自己 製造業 言語 計算機 記述 試験 課題 講義 進捗管理 運用 選択 金融 長期	49.84%
1	報告書：2015年10月	解析テスト用のグループ	2017-06-30 21:13:46	css erp html it javascript sap web アプリケーション ウェブ グループ グループワーク コミュニケーション システム セオリー ゼミ データベース ドイツ ビジネス ビジネスシステム プレゼンテーション プログラミング プロジェクト メンバー リーダーシップ 中国語 京都情報大学院大学 企業 作成 使用 内容 出題 勇矢 友人 受講 報告 大学院 学習 実施 山中 形式 情報処理試験 操作 期間 来季 業務 概論 機会 活用 演習 状況 理事長 理解 生活 知的財産権 知識 研修 紹介 経過 統合 総合職 総評 自己 英語 言語 計算機 課題 調査 講義 運用 選択 長期 非常	45.08%

図 1　「報告書：2015 年 12 月」の分析結果表示画面

ルーム名「報告書：2015年12月」の構成要素のスコアリング

名詞						動詞					形容詞					
順位	単語	出現回数	グラフ	スコア	グラフ	順位	単語	出現回数	グラフ	スコア	グラフ	順位	単語	出現回数	グラフ	スコア	グラフ
1	システム	8		3.48		1	学ぶ	7		8.75		1	赤い	1		14.29	
2	知的財産権	3		3.03		2	進める	4		5		2	幅広い	1		10	
3	ビジネス	6		2.61		3	用いる	2		2.5		3	長い	1		7.59	
4	経営資源	3		2.61		4	続く	3		1.71		4	大きい	2		6.5	
5	html	3		2.37		5	務める	2		1.5		5	多い	1		1.6	
6	中間テスト	2		2.02		6	関わる	1		1.25		6	ない	1		0.07	
7	ビッグデータ	2		2.02		7	培う	1		1.25		7	-	-		-	
8	情報処理試験	2		2.02		8	掘り下げる	1		1.25		8	-	-		-	
9	グループワーク	2		2.02		9	生かす	1		1.25		9	-	-		-	
10	javascript	2		1.83		10	表す	1		1.25		10	-	-		-	
11	ゼミ	8		1.74		11	迷う	1		1.25		11	-	-		-	
12	期末試験	2		1.74		12	図る	1		1.25		12	-	-		-	
13	講義	8		1.56		13	伴う	1		1.25		13	-	-		-	
14	分野	6		1.3		14	取り扱う	1		1.25		14	-	-		-	
15	erp	2		1.25		15	取り入れる	1		1.25		15	-	-		-	
16	形式	5		1.09		16	受ける	2		0.91		16	-	-		-	
17	言語	5		1.09		17	教える	1		0.75		17	-	-		-	
18	リーダーシップ	2		1.01		18	つける	1		0.75		18	-	-		-	
19	コミュニケーション	2		1.01		19	向ける	1		0.63		19	-	-		-	
20	アプリケーション	2		1.01		20	加える	1		0.62		20	-	-		-	
21	プレゼンテーション	2		1.01		21	変わる	1		0.56		21	-	-		-	
22	マネジメント	2		1.01		22	呼ぶ	1		0.5		22	-	-		-	
23	スクリプト言語	1		1.01		23	考える	1		0.27		23	-	-		-	
24	グループディスカッション	1		1.01		24	行う	1		0.21		24	-	-		-	
25	ビジネスモデル	1		1.01		25	なる	3		0.19		25	-	-		-	
26	京都情報大学院大学	1		1.01		26	ある	2		0.11		26	-	-		-	
27	ビジネスシステム	1		1.01		27	する	36		0.05		27	-	-		-	
28	データサイエンス	1		1.01		28	できる	1		0.01		28	-	-		-	
29	マークアップ	1		1.01		29	-	-		-		29	-	-		-	
30	スタイルシート	1		1.01		30	-	-		-		30	-	-		-	

図 2　「報告書：2015 年 12 月」のスコアリング表示画面