Vol.2014-AVM-84 No.7
2014/2/21

日本語 WordNet を利用した単語間の類似度計算による画像検索システムに関する研究

陳　豊[†]　　亀山　渉[†]

地域観光振興のため，Web や SNS から地域に関する人気キーワードをリアルタイムに抽出し，それにマッチした番組コンテンツを検索・自動配信するシステムを筆者らは提案している．本報告では，提案システム中のコンテンツ検索に関する部分を報告する．具体的には，HBase と日本語 WordNet を利用した人気キーワードとコンテンツメタデータの単語間類似度計算による画像検索システムについて報告するとともに，作成したシステムに新聞社が実際に使用した写真画像を投入して実験を行った結果を報告する。

Image Retrieval System based on Word-Similarity using Japanese WordNet

Feng CHEN[†]　　Wataru KAMEYAMA[†]

We have proposed a regional information-platform which delivers audiovisual contents automatically that are retrieved with some popular keywords automatically extracted from Web pages and/or SNS related to the region. In this paper, the content retrieval part of the proposed system is described. The image retrieval method based on word-similarity between popular keywords and content metadata using HBase and Japanese WordNet, as well as the experiments using images provided by a newspaper company is explained.

1. まえがき

地方振興の一環として，Web や SNS から抽出した人気キーワードを用いて番組コンテンツを検索・配信するシステム [1]を実現するため，HBase マッピングによる日本語 WordNet を利用した検索手法を筆者らは提案している [2) 3)]．本報告では，新聞社が実際に紙面に使用した写真画像を用いた検索システムの実装とその評価について報告する．具体的には，日本語 WordNet を利用して写真画像メタデータの上位語を検索し，Wu-Palmer[4)]アルゴリズムによって人気キーワードとの類似度を HBase で高速に計算する手法について報告する．

2. 提案手法と動作

まず，すべての画像データのメタデータを抽出し，ファイル名などの情報と一緒に MySQL に保存する．これにより，各画像に唯一の ID を割り当てる．次に，画像のメタデータから日本語 WordNet を用いて上位概念を抽出し，画像情報と併せて HBase データベースに保存する．人気キーワードに対応する画像を検索する際には，人気キーワードから日本語 WordNet を用いて上位概念を抽出し，それを先のデータベースと照合する．画像メタデータと人気キーワードが同じ上位概念を共有している場合は，両者の意味が近いと考えられる．そこで，同じ概念を共有する人気キーワードと画像のメタデータの類似度を WuPalmer アルゴリズムで計算する．類似度計算結果が閾値より大きい場合，その画像を人気キーワードに対する検索結果として採用する．提案システムの動作フローを図 1 に示す．

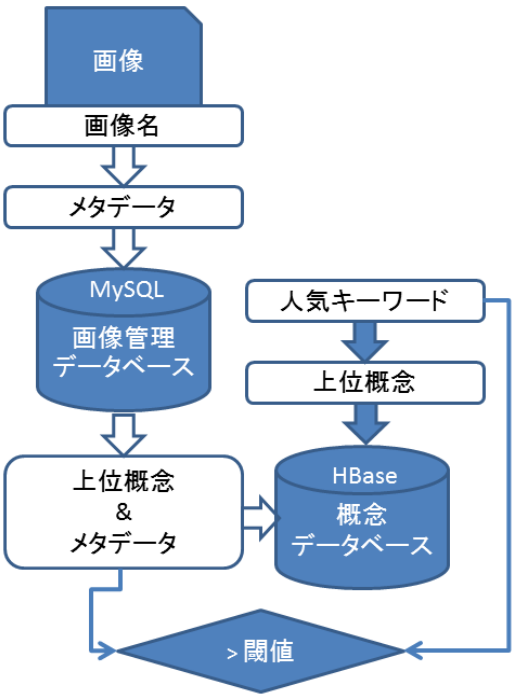

図 1　システム動作フロー

† 早稲田大学　大学院国際情報通信研究科
GITS, Waseda University

3. 画像データ投入

3.1 長野日報社の新聞写真データ

本システムでは，長野日報社が実際に使用した 2010 年 1 月から 2013 年 10 月までの新聞写真データ（125,720 枚，合計 116.85GB）を使用した．この写真画像データには、説明のための短い文書がファイル名としてつけられている．ファイル名は，最初の 2 文字で地域名と対応する記事の分類を表し，全角スペース/半角スペース/：等がそれに続き，その後に写真の説明文書が続いている．画像はすべて JPG ファイルである．2012 年 10 月 6 日のファイルの一部を図 2 に示す．

図 2　2012 年 10 月 6 日の写真画像データ（一部）

3.2 メタデータ抽出

3.1 で述べた写真画像の説明文書からメタデータとして利用できる単語を抽出するため，形態素解析ツールを使用した．具体的には，京都大学情報学研究科と NTT コミュニケーション科学基礎研究所との共同研究を通じて開発されたオープンソース形態素解析エンジンである Mecab[5]を使用した．また，形態素分析ツールとして Igo[6]を使用した．

表 1　形態素解析処理結果例

単語	分析結果	位置
外食	名詞,サ変接続,*,*,*,*,外食,ガイショク,ガイショク	3
の	助詞,連体化,*,*,*,*,の,ノ,ノ	5
ススメ	名詞,一般,*,*,*,*,*	6
３	名詞,数,*,*,*,*,３,サン,サン	9
０	名詞,数,*,*,*,*,０,ゼロ,ゼロ	10
膳	名詞,接尾,助数詞,*,*,*,膳,ゼン,ゼン	11
	記号,空白,*,*,*,*,　,　,	12
トマト	名詞,一般,*,*,*,*,トマト,トマト,トマト	13
ジュース	名詞,一般,*,*,*,*,ジュース,ジュース,ジュース	16
うどん	名詞,一般,*,*,*,*,うどん,ウドン,ウドン	20

2012 年 10 月 6 日の画像ファイルの一つである「下共　外食のススメ３０膳　トマトジュースうどん.JPG」の処理結果を表 1 に示す．本システムでは，メタデータとして名詞だけを採用する．その理由の一つとして，日本語 WordNet で一番多く上位下位関係を持っている品詞であることがあげられる．形態素解析の結果は表 2 に示す MySQL のテーブルに保存する．ここで，id には MySQL が自動的にユニークな番号を割り振る．name には画像ファイル名，path に画像ファイルのパス名，prefix に写真の地域名，data_date に画像ファイルの日付を保存する．そして，keywords に形態素解析した結果抽出された名詞をメタデータとしてカンマで区切った文字列で保存する．

表 2　MySQL テーブル

Field	Type	Null	Key	Default
Id	int(11)	NO	PRI	NULL
name	varchar(1024)	NO		NULL
path	varchar(1024)	NO		NULL
prefix	varchar(64)	YES		NULL
keywords	varchar(512)	NO		NULL
data_date	date	NO		NULL

3.3 HBase に保存

抽出されたメタデータについては，日本語 WordNet から所属概念と上位概念を検索し，画像情報と一緒に HBase の syn_con テーブルに保存する．利用する HBase テーブルは文献 3)で発表したものを修正して用いている．個々のテーブルのスキーマを表 3 に表す．

表 3　HBase テーブル設計

Table	Row Key	Column	Value
lookup	word_id	lang：word_id	word
word_syn	word_id	forward-distance：synset_id	synset_exp
syn_word	synset_id	lang：word_id	rank_word
syn_con	synset_id	backward-distance：content_id	metadata_filepath

まず，MySQL の keywords 中のカンマで区切られた文字列から，個々のメタデータを抽出する．次に，メタデータが lookup テーブルに存在しているかどうかをチェックする．存在した場合，そのメタデータの word_id を取得する．この word_id で word_syn テーブルから直接所属概念もしくは上位概念を検索する．

検索された所属概念もしくは上位概念の synset_id は syn_con テーブルの Row Key として保存する．Column には文献 3)で述べた拡張性のため，Column Family に，メタデータから直接所属している概念には backward-distance を 1，直接所属していない上位概念には backward-distance を 2 として保存する．文献 3)からの変更点は value で，現在処理しているメタデータと画像ファイルパス名を保存する．これは，画像ファイルのどのメタデータと類似度を計算するのかに用いられる．

4. 検索

4.1 検索手順

人気キーワード A で検索する場合，まず，lookup テーブルからキーワード A の word_id を検索する．この word_id で word_syn テーブルから直接所属概念 a，もしくは上位概念 b を検索する．次に，概念 a もしくは概念 b で，syn_con テーブルから関連している画像ファイルを検索する．また value には画像ファイルのメタデータを保存していることから，キーワード A とメタデータの類似度を Wu-Palmer アルゴリズムで計算する．計算した類似度が閾値より大きい場合，それを検索結果として採用する．最後は，検索条件（類似度順，日付順等）でソートした結果を出力する．

4.2 検索結果

本研究では，検索結果の類似度は Wu-Palmer アルゴリズムで計算している．Wu-Palmer アルゴリズムは 2 つの単語の類似度を計測するために，単語の知識構造上の深さ及び共有する最上の上位概念を利用する手法である [7]．Wu-Palmer アルゴリズムで計算した類似度の値の範囲は 0～1 となる．

5. 実験

今回の実験では，正確な比較のために，Linux サーバ 1 台で提案システムを評価した．実験環境を表 4 に示す．

表 4　実験環境

CPU	Intel(R) Core i7 3.33GHz
Memory	12GB
OS	CentOS 6.3
Java	Java 1.7
HBase	Hadoop 2.0-CDH 4.1.2，Hbase 0.92-CDH 4.1.2, Pseudo-Distributed
SQLite	Package：org.sqlite.JDBC

5.1 類似度順の検索

閾値に 0.7 を設定し，5 回検索した処理時間の平均値を求め，類似度が最も高い 5 つの検索結果を採用した．ここでは，検索されたメタデータの重複を許している．「リンゴ」の直接所属概念を利用した検索結果を表 5 に，上位概念を利用した検索結果を表 6 に示す．「酒」の直接所属概念を利用した検索結果を表 7 に，上位概念を利用した検索結果を表 8 に示す．

表 5　「リンゴ」の直接所属概念を利用した検索結果（重複あり）

順番	処理時間	類似度	メタデータ
1	868ms	1	りんご
2	872ms	1	りんご
3	857ms	1	りんご
4	865ms	1	りんご
5	851ms	1	りんご
平均	862.6ms	1	-

表 6　「リンゴ」の上位概念を利用した検索結果（重複あり）

順番	処理時間	類似度	メタデータ
1	1084ms	1	りんご
2	1062ms	1	りんご
3	1061ms	1	りんご
4	1074ms	1	りんご
5	1076ms	1	りんご
平均	1071.4ms	1	-

表 7　「酒」の直接所属概念を利用した検索結果（重複あり）

順番	処理時間	類似度	メタデータ
1	987ms	0.9231	アルコール
2	958ms	0.9231	アルコール
3	958ms	0.9231	アルコール
4	971ms	0.9231	アルコール
5	920ms	0.8333	焼酎
平均	958.8ms	0.90514	-

表 8　「酒」の上位概念を利用した検索結果（重複あり）

順番	処理時間	類似度	メタデータ
1	1059ms	1	コーヒー
2	1028ms	1	カフェ
3	1022ms	1	カフェ
4	1035ms	1	ワイン
5	1017ms	1	ワイン
平均	1032.2ms	1	-

上位概念を利用した処理時間が，直接所属概念を利用した処理時間より長くなっている．また，表 7 の類似度から，同じ概念に直接所属していても，類似度が 1 より小さい場合もある．

上位概念を使用した検索結果の表 6 と表 8 から見ると，「リンゴ」とマッチするメタデータの検索結果はどれも正しいものとなっている．一方，「酒」とマッチするメタデータの検索結果では，類似度は非常に高いが，実際の検索意図には必ずしも当てはまらない．これは，「リンゴ」が日本語 WordNet の中で「酒」より階層が深い単語であり，Wu-Palmer アルゴリズムが単語の知識構造上の深さを考慮しているため，階層が浅い単語に対しては，類似度を大きめに計算してしまうことによると考えられる．

5.2 検索されたメタデータの分析

Wu-Palmer アルゴリズムを使用し，どんなメタデータを検索できるのかを把握するため，検索されたメタデータの重複を許さずに，類似度順で確認する必要がある．

そのため，閾値に 0.7 を設定し，類似度が最も高い 5 つの異なったメタデータの検索結果を分析した．5-1 節と同様に，直接所属概念と上位概念を利用した検索手法で比較を行った．「リンゴ」の直接所属概念を利用した検索結果を表 9 に，上位概念を利用した検索結果を表 10 に示す．「酒」の直接所属概念を利用した検索結果を表 11 に，上位概念を利用した検索結果を表 12 に示す．

表 9 「リンゴ」の直接所属概念を利用した検索結果（重複なし）

順番	類似度	メタデータ
1	1	りんご
2	1	林檎
平均	1	-

表 10 「リンゴ」の上位概念を利用した検索結果（重複なし）

順番	類似度	メタデータ
1	1	りんご
2	1	林檎
3	0.9524	実
4	0.9524	木の実
5	0.8889	プラム
平均	0.95874	-

表 11 「酒」の直接所属概念を利用した検索結果（重複なし）

順番	類似度	メタデータ
1	0.9231	アルコール
2	0.8333	焼酎
平均	0.8782	-

表 12 「酒」の上位概念を利用した検索結果（重複なし）

順番	類似度	メタデータ
1	1	牛乳
2	1	コーヒー
3	1	ティー
4	1	ミルク
5	1	ワイン
平均	1	-

直接所属概念を利用した検索の場合，検索されたメタデータ数が上位概念を利用した場合よりも少なくなっている．階層が深い「リンゴ」の場合も，階層が浅い「酒」の場合も，直接所属概念を利用した検索では 2 つのメタデータが検索されたのみである．よって，上位概念を使用する検索の場合の方が，より広くメタデータを検索できることが分かった．

また，表 10 の上位概念を利用した「リンゴ」の検索結果では，類似度が大きいメタデータとして「実」や「木の実」が検索されていることから，Wu-Palmer アルゴリズムの特徴がよく表れていると言える．一方，表 12 の上位概念を利用した「酒」の検索結果では，意味的に遠いメタデータが多く検索される結果となった．

5.3 日付順の検索

画像データの日付が新しいものを優先的に検索する場合を分析する．

閾値に 0.7 を設定し，日付の新しい順に 5 つの画像検索結果を求めた．「リンゴ」の直接所属概念を利用した検索結果を表 13 に，上位概念を利用した検索結果を表 14 に示す．「酒」の直接所属概念を利用した検索結果を表 15 に，上位概念を利用した検索結果を表 16 に示す．

表 13 「リンゴ」の直接所属概念を利用した検索結果（日付順）

順番	処理時間	類似度	画像ファイル名 日付
1	871ms	1	諏共りんごの会マレット.JPG 2013.10.08
2	858ms	1	辰Ｌ地域活動支援センターりんごオーナー作業.JPG 2013.10.05
3	855ms	1	中共●飲むゼリー「完熟りんごジュレ」を開発 中川の富永農園.JPG 2013.06.22
4	870ms	1	駒共◎６月８日に「奇跡のりんご」木村さん講演会.JPG 2013.05.24
5	864ms	1	諏Ｌ　第５回太田屋りんごの会杯マレットゴルフ大会.JPG 2013.05.13
平均	863.6ms	1	-

表 14 「リンゴ」の上位概念を利用した検索結果（日付順）

順番	処理時間	類似度	画像ファイル名 日付
1	1132ms	0.9524	茅共◎金子美那実さんがおもちゃインストラクター.JPG 2013.10.16
2	1052ms	0.9524	諏共諏訪実図書委員会カンボジアへ絵本を.JPG 2013.10.09
3	1037ms	1	諏共りんごの会マレット.JPG 2013.10.08
4	1043ms	0.8889	諏共　マルメロ収穫２.JPG 2013.10.08
5	1051ms	0.8889	諏共　マルメロ収穫１.JPG 2013.10.08
平均	1063ms	0.93652	-

表 13，14，15，16 から，直接所属概念を利用した検索処理時間においては階層が深い「リンゴ」のほうが短く，上位概念を利用した検索の処理時間においては階層が深い「リンゴ」のほうが長くなっている．同様の傾向は，表 5，6，7，8 にも見られる．一般に，階層が深い単語ほど意味的により狭い概念に分けられているので，同じ概念に所属する同意語が少なく，検索も早いことが考えられる．逆に，階層が浅い単語ほど，意味的により広い概念に所属する可能性が高いので，同意語が多く，処理時間が長くなると考えられる．

表 15 「酒」の直接所属概念を利用した検索結果（日付順）

順番	処理時間	類似度	画像ファイル名 日付
1	984ms	0.8333	南Ｌ　いも焼酎南箕輪会が活動始動.JPG 2013.05.31
2	977ms	0.9231	東共　キリン　ノンアルコール÷チューハイ　ゼロハイ　シャルドネ.JPG 2013.05.30
3	919ms	0.8333	伊Ｌ「西春近いも焼酎の会」，試飲会.JPG 2013.02.22
4	977ms	0.8333	南Ｌいも焼酎南箕輪会始飲会.JPG 2012.12.26
5	977ms	0.8333	茅共　両久保エコファーム収穫祭，芋焼酎で乾杯.JPG 2012.12.01
平均	966.8ms	0.85126	-

表 16 「酒」の上位概念を利用した検索結果（日付順）

順番	処理時間	類似度	画像ファイル名 日付
1	1025ms	1	岡共　エーピーエヌが諏訪地方の観光を印刷したコーヒー発売.JPG 2013.10.15
2	1011ms	1	諏共アートカフェ諏訪塾.JPG 2013.10.14
3	1042ms	1	住宅模型による家づくりカフェ.JPG 2013.10.12
4	1015ms	1	宮Ｌワインセミナー収穫体験.JPG 2013.10.06
5	1047ms	1	宮共宮田産ワイン仕込み始まる.JPG 2013.10.03
平均	1028ms	1	-

6. まとめと今後の課題

本報告では，提案する日本語 WordNet を利用した単語間の類似度計算による画像検索システムに，実際の新聞社の写真データを導入し，Wu-Palmer アルゴリズムを使った検索結果を評価した．検索結果は，検索時間及び精度の観点からはおおむね良好であったと考えられるが，階層が浅い単語では必ずしも Wu-Palmer アルゴリズムがうまく働かない場合もあることが分かった．

今後の課題としては，他の単語間類似度アルゴリズムと Wu-Palmer アルゴリズムの検索結果と比較することがあげられる．また，構築したシステムをモジュール化し，実システムとして運用することを考えたい．

謝辞

本研究の一部は，総務省「戦略的情報通信研究開発推進制度（SCOPE）」（採用課題番号：122304003）の研究助成によるものである．また，長野日報社には貴重な画像データを提供頂いた．ここに記して謝意を表す．

参考文献

1) 三代沢，広瀬，土屋，亀山，小柳，山本，唐澤（英安），唐澤（英長），増沢，“地域情報プラットフォームと通信放送連携システムの開発と評価　～地域観光振興と防災に向けて～”，映像情報メディア学会，放送技術研究会，2013-12-AVM-CS-IE-BCT，BCT25，2013 年 12 月
2) 陳豊，PaoSriprasertsuk，亀山渉，“WordNet を利用した番組等検索システムに関する研究”，電子情報通信学会，2013 年総合大会，D-5-5，2012 年 3 月
3) 陳豊，PaoSriprasertsuk，亀山渉，“HBase マッピングによる日本語 WordNet を利用した番組等検索システムに関する研究”，第 12 回情報科学技術フォーラム FIT2013，D-026，2013 年 9 月
4) Zhibiao Wu，Martha Palmer，“Verbs semantics and lexical selection”，In Proceedings of the 32nd annual meeting on Association for Computational Linguistics (ACL-94)，pp.133-138，1994 年
5) MeCab: Yet Another Part-of-Speech and Morphological Analyzer，http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html，2014 年 1 月 8 日最終確認
6) Igo，http://igo.sourceforge.jp，2014 年 1 月 8 日最終確認
7) 横手健一，ボレガラ・ダヌシカ，石塚 満，“テキスト含意認識に有効な意味類似度変換及びその獲得法”，人工知能学会誌，Vol.28，No.2，pp.220-229，2013 年 2 月