Vol.2014-DPS-159 No.14
Vol.2014-MBL-71 No.14
2014/5/15

特徴語を用いた観光レコメンデーションシステムの検討

小川 開[†1] 杉本 祐介[†1]
内藤 克浩[†2] 菱田 隆彰[†2] 水野 忠則[†2]

近年、スマートフォン, タブレット型端末の登場により、人々の生活をもとにした、巨大なデータが蓄積可能となってきている。また、口コミ情報の投稿や位置情報の取得も容易である。しかし、現在存在する一般的なサービスでは、口コミ情報そのものを掲示し利用している。口コミ情報を解析することで、レコメンドに活かせるのではないかと考え、本研究の目的である特徴語を抽出しレコメンドに使用する方式を提案する。口コミ情報から特徴語を抽出することで、観光地毎の特徴が見えてくる。そのため、ある観光地特有の要素の発見につながり、今までユーザの目にあまり映らなかった、観光地にもスポットが当てられるのではないかと考える。結果過疎化の進んでいる観光地の救済にもつながるのではないかと考え、本研究を提案する。

Basic design of a sightseeing recommendation system using Characteristic Words

KAI OGAWA[†1] YUSUKE SUGIMOTO[†1]
KATSUHIRO NAITO[†2] HISHIDA TAKAAKI[†2] TADANORI MIZUNO[†2]

1. はじめに

近年、スマートフォン、タブレット型端末の登場から、人々の生活にITが身近なものとなった[1]。IT普及率増加[2]に伴い、人々の生活をもとにした巨大なデータの蓄積が可能となってきている。それらのデータをもとにし、レコメンデーションを組み込んだ多種多様なシステムが登場している。それに伴い、位置情報を含んだ口コミ情報の投稿も容易になっている。以前は位置情報を取得するには、専用の機器を使用しなければならなかったが、現在は基本的にスマートフォンやタブレットに搭載されている。そこで、口コミ情報と位置情報から、レコメンドに活用できるのではないかと考え、本研究を進めている。

本研究の口コミ情報解析と同じく、文字列解析の関連研究として、「Twitter分析に基づく位置依存文字列の抽出」[3]がある。この研究では、位置情報付きツイート 50 万件の中から、位置依存性の高い文字列を抽出する手法について、記述されている。初めに、Yahoo!日本語形態素解析APIとYahoo!キーフレーズ抽出APIをも用いてキーワードを抽出している。すべてのキーワードについて、出現回数を計測し、緯度経度についての標準偏差を測定する。その中から、標準偏差の値が小さいもの、すなわち位置情報と関わりが高い、Twitterコメントを検出、計測することが可能となっている。

また、「ビッグデータ処理の展望-変ぼうするデータ分析技術の動向-」[4]では、ビッグデータの一つである、大規模テキスト・データの分析について記述してある。大規模テキスト・データを分析し、社会の影響とどのような関わりがあるのか。テキスト分析の対象をソーシャル・メディアとしているのだが､情報共有で消費者の購買意欲の変化、震災後の物品不足との関わりなどを検証している。また、単にテキスト・データのコンテンツ量を見ているのではなく、時間空間的属性、人やコミュニティのネットワーク構造、メディアの持つ双方向性など複雑な属性を理解した上での分析が必要だと述べている。

特徴語の抽出としては、「テキストコーパスにおける特徴語抽出のための分析ツール」[5]では、形態素情報を手がかりに抽出している。形態素情報から、複合名詞を順序付き単語リストとみなして分析する方法を記載してある。この分析ツールでは、文の関わりの強さ「結合度」、情報量の観点から「出現度」、ある語と直前の別の語で新たな語を構成しているか判別する「前接度」、反対に直後の言葉で「後

†1 愛知工業大学大学院経営情報科学研究科
Graduate School of Business Administration and Computer Science, Aichi Institute of Technology,

†2 愛知工業大学情報科学部情報科学科
Facutly of information Science
Aichi Institure of Technology

Vol.2014-DPS-159 No.14
Vol.2014-MBL-71 No.14
2014/5/15

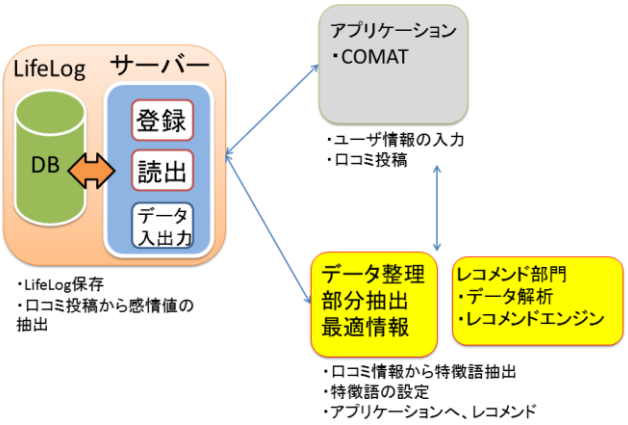

図1:AHLEの概要図

接度」、文脈とどのくらい関わりがあるか判別する「文脈度」、以上5点を総合的に判断する「重要度」の6つの点から判断し特徴語を抽出している。結果、セグメント集合を切り出し、木構造表現に変換し計算を行い、各
尺度に基づく特徴語のランキングを閲覧可能にしている。

一般的に、現在普及しているレコメンドシステムの多くは、ユーザの入力データを元に、システム開発者が設定した基準に基づいてレコメンドを行う。設定した基準に変更や更新がなければ一定の出力を行うシステムとなっている。条件があっていればどのような地域、どのような要素でも使用できるのだが、トレンドなどの流行り物、人口の過疎過密、情報量の過不足等地域の特色に対応することは難しい。各地域には各地域の特色が存在しており、地域ごとに対応したレコメンドシステムを開発すべきだと考える。

そのため現在、一般的なサービスで行われている口コミ情報の提示だけでなく、更に口コミ情報や位置情報を活かせないかと考えた。単に提示するだけではなく、口コミ情報から、特徴のある言葉、特徴語を抽出しレコメンドに活用できると考え提案する。特徴語を抽出するにあたって、まず、どの単語を特徴語として抽出すればよいのか。また、抽出した特徴語をどのように使用すれば、レコメンドに活用できるのか、更に抽出した特徴語をどのように分類するのか、など様々な問題がある。本研究では現段階での特徴語の使用方法を記述する。

また、本研究は「口コミデータを活用するデータベースシステムの実現」[6]の解析エンジンを担当している。上記のデータベースではロコミを行えるアプリケーションを対象とし、口コミ情報を解析、数値化し有効活用を考えている。口コミ情報の対応アプリケーションとして、農産活用型観光誘導アプリケーション COMAT(Citizens cooperation MApping for Toyota)[7]において、観光地のレコメンドを行うことにした。以下 2 章にシステム全体の概要、3 章レコメンドエンジンの説明、4章で特徴語を用いたレコメンド、考察、5章で本研究のまとめを記載する。

2. システムの構成

システムの構成として、LifeLog やロコミから感情値解析を行う DB、AHLE、農産活用型観光誘導アプリケーション COMAT、本研究が担当する解析エンジン、レコメンド部門からなる。以下 2.1 節で AHLE 全体について、2.2 節でアプリケーションの一例として、COMAT について、2.3 節データ整理、レコメンド部門について、それらの説明を記載する。

2.1 AHLE

観光アプリケーション COMAT、それら Lifelog を保存し、ロコミ情報からキーワード抽出感情解析を行う DB、DB の情報を解析、分析し COMAT のユーザへ観光地、農産地のレコメンドを行う解析エンジンから成り立っている。図 1 に AHLE 全体概要図を記載する。

DB は「ロコミデータを活用するデータベースシステムの実現」[6]を使用している。OS は CentOS を使用している。また、Apache と MySQL で管理している。形態素解析を行い、プログラムで集計し後に計測を行っている。システム全体の研究目的は、ロコミ情報から感情値を取得、それらを解析、使用しユーザへレコメンドすることである。今回は本システムに適するアプリケーションの 1 例としてスマートフォンアプリケーション COMAT を採用している。COMAT についての説明は下記 2.2 節に記載する。

2.2 COMAT

COMAT は農産活用型観光誘導アプリケーションである。既存の観光アプリケーションとは違い、ユーザに観光地情報、農産物直売所情報を提示することで、観光地に訪れた人は農産物直売所を、農産物直売所を訪れた人には観光地情報が得られる。観光地と農産物直売所を効果的に提示する事によりユーザの行動範囲の増加と新たな観光場所の認知を狙うといった特徴がある。利用形態を以下に記載する。利用手順としてユーザ登録を行い、ユーザを観光地へ誘導し訪問させる。そして、ユーザに観光地へのロコミ情報の投稿を促す。恩典として豆知識を提供し、観光場所の認知向上、観光客増加が狙いである。また COMAT では登録時のユーザ情報（性別、年齢、名前）、観光地へのロコミ情報（ロコミ投稿時の日付、時間、ロコミ情報の種類、ロコミ場所の位置情報、投稿された端末名、投稿画像、評価値）、観光地情報(観光地の緯度経度、観光地の詳細情報)が登録される。表 1 にロコミ情報登録時の XML データ例を記載する。

Vol.2014-DPS-159 No.14
Vol.2014-MBL-71 No.14
2014/5/15

表 1:口コミ情報入力データ

	タグ	値
	lifelog	一件分のユーザー情報と口コミ情報(下記タグを内包)
ユーザー情報	person	ユーザーの性別、年齢
	name	ユーザーの名前
口コミ情報	date	日付
	time	時間
	category	口コミ情報の種類
	location	口コミ情報の場所情報
	geometry	口コミ情報の緯度、経度、精度
	place	口コミ場所名
	device	端末のデバイス種類
	content	口コミ内容
	path	画像パス
	evaluation	評価値

2.3 データ整理、レコメンド部門

データ整理、レコメンド部門ではアプリケーションでの取得情報、口コミ情報、またDBで口コミ情報における特徴語を抽出したデータを使用しレコメンドを行う。今回採用しているアプリケーションCOMATは農産活用型観光誘導アプリケーションであるため、観光地へのレコメンド、地域ならではの口コミ情報の特色を活かしたレコメンドを行わなくてはならない。今回はレコメンドシステムについての研究を述べる。今回搭載するアプリケーションCOMATは愛知県豊田市での稼働を予定している。そのため、豊田市ならではの地域性、口コミ成功を使用し地域の特色に反映させることを設計基準した。結果、他のユーザ情報によってレコメンドを行う協調フィルタリング型のレコメンドシステム「アソシエーション分析」を採用した。アソシエーション分析を使用するに当たって、データがないとレコメンドが行えない。そのため初期データを稼動前に入力しておき、稼動後にデータを取得していく。

3. レコメンドエンジン

3.1 レコメンドアルゴリズム

レコメンドのアルゴリズムには R 言語、アソシエーション分析を採用した。R 言語とは統計計算とグラフィックスの為のシステムである。また、R 言語は、1980 年代にデザインされそれ以来統計関係者の間で広く使われてきた S 言語の方言である。言語の構文は C 言語と外見上の類似性を持つ。また“言語による計算”を可能にし、それにより表現式を入力に取る関数を書くことが可能である。[8]
本研究ではアンドロイドアプリケーションのユーザを想定している。アンドロイド端末からR言語を操作することは難しいと考え、入力をHTMLベース、出力をXML形式で行う。またユーザの条件に基づきレコメンド可能な手法を考え、DBから条件があったものを出力する、XML Extractプリプロセッサ、XML Extractプリプロセッサの出力をR言語で使用できる形式に変換を行うJRコンバータを通して処理を行う。レコメンド要求時に行う処理を図2

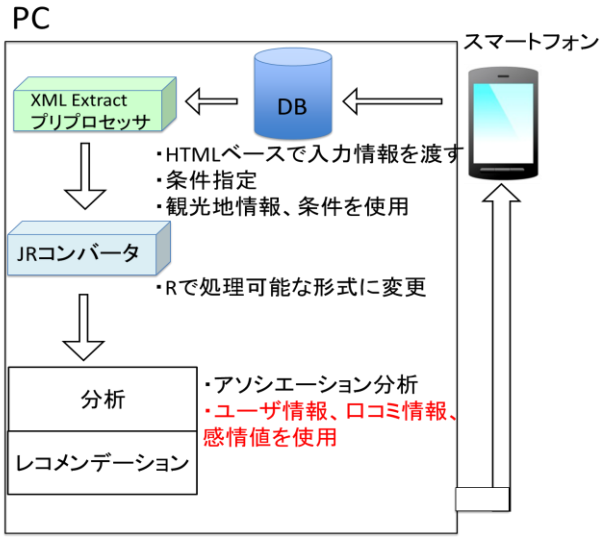

図2:レコメンド要求

に記載する。また、レコメンド要求時、各項目の説明をいかに行う。

◆ XML Extract プリプロセッサ

ユーザのアプリケーションから、ユーザの条件を受け取り DB から、要求に適した観光地のデータのみを抽出してくる。XML 形式での出力となるため、この処理のみでは R 言語で使用できない。表 2 に XML Extract プリプロセッサで処理を行う、観光地情報。図 3 に XML Extract プリプロセッサ出力例を記載する。

◆ JR コンバータ

XML 形式の観光地データを R 言語で使用できる list 形式に変換する。また、出力されたデータを R 言語に転送し、処理を行わせる。

3.2 使用パラメータ

レコメンドで使用するパラメータとして、ユーザ情報を他のユーザ、観光地情報、条件、もしくは両方をと比較し協調フィルタリングを行うもの。ユーザが選択した条件を観光地の条件でフィルタリングし省くもの。口コミ情報から観光地情報、キーワード抽出、感情を抽出しレコメンドに使用するもの、以上3点に分けて説明する。

◆ ユーザ情報:

他のユーザと比較し、類似のユーザ情報をレコメンド行う。使用パラメータとしては、ユーザの年齢、性別に加え、口コミ情報から観光地に訪れた際の同行者の人数、同行者との間柄、観光地への目的などを抽出する。以上の情報を使用し他のユーザ情報とで強調フィルタリングで観光地のレコメンドを行う。

Vol.2014-DPS-159 No.14
Vol.2014-MBL-71 No.14
2014/5/15

表2: 観光地情報

タグ	値
name	場所名
geometry	緯度、経度
address	住所
introduction	場所への情報
path	画像ファイルへのパス
parking	駐車場情報
coment	場所へのコメント
emotion	感情値

```
<?xml version="1.0" encoding="UTF-8"?>
<lifelog>
    <person type="user" gender="male" age="20">
        <name>testuser</name>
        <address type="present">愛知県豊田市</address>
        <address type="birth">愛知県豊田市</address>
    </person>
    <content>豊田市役所なう</content>

    <evaluation type="ovaeall" value="4.5" />

    <emotion type="joy" value="5.0" />
    <emotion type="trust" value="4.5" />
    <emotion type="fear" value="4.0" />
    <emotion type="surprise" value="3.5" />
    <emotion type="sadness" value="3.0" />
    <emotion type="disguest" value="2.5" />
    <emotion type="anger" value="2.0" />
    <emotion type="anticipation" value="1.5" />

</lifelog>
```

図3: XML Extractプリプロセッサ出力例

◆ 観光地情報

ユーザの要望を設定し、フィルタリングをかけユーザに適した情報のみを選出させる。使用パラメータとしては、DB内の場所情報に記載した、現在地から目的地への移動時間、出発地からの移動手段、目的地情報、駐車場情報である。これらの情報を加味しユーザ要望があった場合、条件に合うもののみをユーザに提供する。

◆ 特徴語

観光地への口コミ情報から、一定のキーワードを含む言葉、地域の特性に係る言葉を抽出する。現段階では検討中であるが、口コミ情報から取得できる単語を使用しレコメンドしてムに使用する。特徴語については 4 章に詳しく記載する。

3.3 レコメンドエンジン

アソシエーション分析とは 巨大なデータベースから、価値あるアソシエーション・ルールを抽出するテータマイニングである。効率的に見つける方法かアソシエーション分析である。具体的には、「商品Aを買うと商品Bも買う」というようなアイテム同士の関わりあいを限なく方式である。また、アソシエーション・ルールが有効を判断する際に、何らかの評価指数が必要である。用いられる指数として、ルール全体の出現率、支持度（support）、アイテム同士の結びつきの強さ、確信度（confidence）、アイテム単体とルール同士の出現率の比率、リフト(lift)があり、これらの数値を使用しシステム設計者はルールの優越を決定する。

4. 口コミ情報を用いたレコメンド

口コミ情報から特徴のある言葉を抽出し、レコメンドに用いる。4.1 節にレコメンド手順、4.2 節に特徴語の判定方法、4.3 節に特徴語のリスト方法、4.4 節に現段階で完成している手順 3 まで、サンプルデータを用いた結果を記載する。

4.1 手順

手順 1: 口コミ情報から単語の抽出

各観光地に対して行われた口コミ情報から、単語を抽出する。その中から意味のない単語を省いき、抽出した単語の回数を計測、各観光地のごとに記録しておく。テキストから単語の検出には EKwords（有限会社 DJ SOFT）[10]を使用した。EKwords とは日本語/英語の文章データからキーワード(単語、連語)を抽出し集計するソフトである。

手順 2: 同意味単語の統一

抽出した単語から、同じ意味を持つ言葉を丸め、特徴語の量を計測する。各口コミ情報から同一の意味を持つが、文字の違う単語をまとめ、ひとつの単語としてカウントする。口コミ投稿を行うユーザによって同一の意味を持つ特徴語を使用していても、文字の違いから、異なる特徴語と判断するのを防ぐためである。今回は、一連の動作を実行することを目的としたため、同意味の単語が出現しないテストデータを使用した。現在同意味の単語の処理は検討中である。候補としては「文脈情報による同義語辞書作成支援ツール(語彙・概念の獲得と同義語)」[10]を使用し、ある程度手動で登録して自動化を考えている。

手順 3:特徴語の判定

各特徴語に対して、観光地の固有の特徴語、ポジティブな要素を持つ特徴語、ネガティブ要素を持つ特徴語の 3 種類に選別する。

Vol.2014-DPS-159 No.14
Vol.2014-MBL-71 No.14
2014/5/15

表 3:口コミ情報から名古屋城の特徴語抽出例

特徴語	言葉
固有の特徴語	城,本丸,復元,改修,桜,武将
ポジティブの特徴語	ビックリ,子供,家族
ネガティブの特徴語	人混み,行列,有料,暑い

手順 4:特徴語のリストを作成

特徴語のリストを作成し、ユーザに提示し、選択してもらう。各項目から

手順 5:特徴語を用いたレコメンド

選択された特徴語から、適した観光地のレコメンドを行う。ただ単に観光地を提示するだけではなく、組となる観光地の提示や、ユーザに適した観光地を含む観光ルートを作成し、ユーザへ提供する。

4.2 特徴語の判定

各特徴語について、各観光地に共通して持たない、出現数が少ない特徴語を固有の特徴語、共通して持つがポジティブな意味を持つ特徴語をポジティブの特徴語、反対にネガティブな意味を持つ特徴語をネガティブの特徴語とし、判別する。表 3 に例としてトリップアドバイザー[11]から名古屋城についての口コミを元に特徴語の判定結果を記載する。

ユーザの趣向を考慮し、目的としたい要素を固有の特徴語、なるべく選択したい要素として、各観光地の持つ印象となる特徴語をポジティブの特徴語、一部ユーザにとっては避けて通りたい要素として、省いておきたい要素をネガティブの特徴語とし判別、選択できるようにした。判断基準として､｢インターネット上の書き込みに含まれる感情についての調査｣ [12]の感情語辞書を使用し、喜び、悲しみ、受容に含まれている語と、連結して使用されている言葉をポジティブと判断した。また、ネガティブも同様に、恐れ、怒り、驚きに含まれている語と連結して使用されている言葉を、ネガティブと判断した。その他記載されていない単語は、手動で判断した。今後参照元以外にも判断する基準が必要となってくるであろう。今後の展開として、登録していない単語を自動的に判断するのではなく、ユーザに判断を任せる仕組みを加えていきたい。

4.3 特徴語のリスト作成

検出された特徴語のリストを掲示し、選択してもらう。選択された特徴語を用いて、関連のある観光地を重要視しユーザに観光地のレコメンドを行う。ユーザが最も重要視

する要素を固有の特徴語から選択してもらう。次に好印象な要素を、ポジティブの特徴語から選択、逆にユーザが

表 4:JR セントラルタワーズ、単語抽出例

言葉	単語 複語	語数	出現数
レストラン	単	1	28
ビル	単	1	19
名古屋	単	1	18
名古屋駅	単	1	18
レストラン街	復	2	13
お店	単	1	10
ハンズ	単	1	9
ホテル	単	1	9
行く	単	1	9
東急	単	1	9
東急ハンズ	復	2	9
JR	単	1	8
24 時間	単	1	8
マリオットアソシアホテル	単	1	7
タワー	単	1	6

好まない要素をネガティブの特徴語から選択してもらうことで、ユーザの考えに近いレコメンドが行えるのではないかと考えている。また、ユーザへの表示方法について、ただリスト形式で提示するのでは効果が薄いと考えている。将来的に、似ている特徴語の丸め、特徴語同士の意味合いを考え、効果的な表示を検討している。

4.4 サンプルデータ用いての実行結果

サンプルデータとして､トリップアドバイザーを使用し、サイト内の「JRセントラルタワーズ」についての74件のレビューを解析した。解析結果の内、出現数が高い上位15件を表4に記載する。なお、単語複語の項目には、単語で構成されているものを単、複語で構成されているものを複と記述してある。表から、出現数が高い単語のうちでも、同意味を持つ単語が多々出現していることがわかる。

次に、同意味を持つ単語をまとめた情報を表5に記載する。同一後にも様々なパターンが存在することが確認できた。単に同一の意味を持つが、レストランやレストラン街といった言葉表現の違いがある同一意味の単語。タカシマヤ、高島屋とった名称は同じだが、入力の違う単語。地理情報から、近くの場所を推定し特定できる単語。今回のテストデータではJRや駅とあるが、JR名古屋駅の周辺情報の口コミから名古屋駅と判別できる。

全く同じ単語を使用しているが、会話の流れから意味合いが全く違うもの。現在はテスト段階のため、すべての文章を読み手動で判別した。将来的には自動化を考えているが、様々なパターンがあるため、すべてを機械にませるの

表5:サンプルデータを用いた同一語丸め例

カテゴリ	同意味を持つ単語	丸めた後	検測方法
同意味	レストラン レストラン街 飲食店	レストラン	同意味の単語で辞書を作成
同名称	タカシマヤ 高島屋	高島屋	異なる読み方で辞書を作成
名称不足、地理情報から推測	駅 名駅 JR	名古屋駅	位置情報から推測
文脈	高い 高級	高級	前後の文脈から推測
文脈	高い 高層ビル	高層	前後の文脈から推測

は難しいと考える。ユーザの力を借りながら、対応できる仕組みを考え、実装していくことを目標にして、開発を行っていく予定である。

5. おわりに

現在特徴語を用いたレコメンドの手段を提案した。しかし、基礎で取り扱う物について決まった程度であり、個々の詳しい仕様を確率できていない。その上で、試行錯誤を行いユーザに適した仕様を検討していきたい。また、特徴語は、場所に対して行う口コミ情報から取得している。そのため、場所によって様々な特徴語が取得可能だと考えることができる。こちらが設定した基準ではなく、特徴語を抽出といった、場所毎の違いを活かしたレコメンドを実現できるシステムを組まなければいけないだろう。地域ごとの特色を活かし、過疎地域の観光地救済に役に立てば幸いだと考えている。

次に、同意味単語の丸めだが、同一意味の辞書作成、位置情報、文脈の判断等、様々な手段を用いないと実現が難しいと今回のテストで理解できた。各要素の組み合わせを考え、システムを作成していかなくてはならない。

今後の展開として、レコメンド方式の確立、同一語丸めシステムの自動化、この2点を重点的に作成していき、システムの全体の完成を考えている。その上で、レコメンドに対してユーザ調査を行っていき、より良いレコメンドシステムの作成を予定している。

参考文献

1) 総務省:平成 24 年通信利用動向調査の結果 ,http://www.soumu.go.jp/johotsusintokei/statistics/data/130614_1.pdf ,(参日:2013/09/17)

2) 総務省 情報通信国際戦略局 情報通信経済室:情報通信 国際戦略局 情報通信経済室ICTインフラの進展が国民のライフスタイルや社会環境等に及ぼした影響と相互関係に関する調査研究報告書,http://www.soumu.go.jp/johotsusintokei/linkdata/h23_06_houkoku.pdf,(参照日:2013/09/17)

3) 荒川豊、田頭茂明、福田晃: Twitter 分析に基づく位置依存文字列の抽出,情報処理学会研究報告. モバイルコンピューティングとユビキタス通信(MBL),vol.2010-MBL-55 No.10,pp.1-6(2010.8)

4) 武田浩一、井手剛:ビッグデータ処理の展望. －変ぼうするデータ分析技術の動向－, PROVISION Winter 2012 No.72 .pp70-75

5) 相澤 彰子:テキストコーパスにおける特徴語抽出のための分析ツール,情報処理学会研究報告. 情報学基礎研究会報告 2001(20), pp.113-120,(2001.03)

6) 杉本祐介、土井千章、中川智尋、太田賢、稲村浩、水野忠則 、菱田隆彰:口コミデータを活用するデータベースシステムの実現, 情報処理学会研究報告,モバイルコンピューティングとユビキタス通信(MBL), vol. 2014-MBL-70 No.44, pp.1-6(2014.3)

7) 水上貴晶、早矢拓也、五十里秀人、菱田隆彰、水野忠則:農産活用型観光誘導アプリケーション COMAT の開発,情報処理学会研究報告. モバイルコンピューティングとユビキタス通信(MBL),vol.2014-MBL-70 No.48,pp.1-8(2014.3)

8) 間瀬茂、東京工業大学情報理工学研究科：R 言語定義 (R Language Definition) Version 1.1.0,http://cran.r-project.org/doc/contrib/manuals-jp/R-lang.jp.v110.pdf(参照日:2014/04/01)

9) EKwords（有限会社 DJSOFT）,http://www.djsoft.co.jp/

10) 寺田昭,吉田稔,中川裕志:文脈情報による同義語辞書作成支援ツール(語彙・概念の獲得と同義語),情報処理学会研究報告,自然言語処理研究会 2006(124), pp87-94, (2006.1)

11) トリップアドバイザー,http://www.tripadvisor.jp/

12) 杉本祐介、菱田隆彰、水野忠則：インターネット上の書き込みに含まれる感情についての調査,情報学ワークショップ 2013(WiNF2013), pp.195-199(2013.12).