地方公共団体のオープンデータ化に向けた webページ間の関係性抽出技法の提案

近藤　拓也† 　村山　敬祐† 　杉山いおり† 　遠藤　雅樹†† 　横山　昌平†††
石川　博†

† 首都大学東京システムデザイン学部　〒 191–0065 東京都日野市旭が丘 6–6
†† 首都大学東京大学院システムデザイン研究科　〒 191–0065 東京都日野市旭が丘 6–6
††† 静岡大学大学院情報学研究科　〒 432–8011 静岡県浜松市中区城北 3–5–1
E-mail: †{kondo-takuya@ed.,murayama-keisuke@ed.,sugiyama-iori@ed.,ishikawa-hiroshi@}tmu.ac.jp,
††endo-masaki@ed.tmu.ac.jp, †††yokoyama@inf.shizuoka.ac.jp

あらまし　各地方公共団体の web ページでは様々な公共データが公開されている．しかし，いくつものリンクを経なければ目的のデータに辿りつけないことや，同種のデータでも扱う団体が異なるため別々の形式にまとめられていることがある．そこで，各地方公共団体の web ページのオープンデータ化，アクセシビリティの向上を図ることで，基幹データの有用性や価値を高める目的で研究を行った．本論文では複数の地方公共団体の web ページを用いて，クラスタリングや分類を行い，関連する情報が記載された web ページをまとめ，またリンクの繋がる web ページ間の類似度を算出した結果について報告する．

キーワード　オープンデータ，地方公共団体，web マイニング，クラスタリング，LOD

1. はじめに

近年，日本の行政機関においてビッグデータ，オープンデータを利活用する動きが高まっている．例えば，総務省では総務省統計局管轄の統計調査の状況を次世代統計利用システム (注1) にまとめ，API を公開している．経済産業省もオープンデータ実践に向け試験的に web ページ (注2) を公開している．その他の省庁でもオープンデータ化を図るため，各々の基幹データを公開している．また，それらを包括すべく内閣官房情報通信技術総合戦略室がデータカタログサイト (注3) を試行版として公開し始めた．地方公共団体でも 2013 年 4 月 1 日に武雄市，千葉市，奈良市，福岡市の 4 市合同で「ビッグデータ・オープンデータ活用推進協議会」を設立し，地方公共団体のビッグデータ，オープンデータ利活用の先駆けとなっている．さらに全国に 1,742 箇所ある地方公共団体 (2014 年 1 月 1 日時点) も各々の基幹データを各 web ページ内にまとめ公開している．しかし，地方公共団体間の PDF，Microsoft® Word ，Excel などのファイルが存在する web ページに関しては，基幹データファイルは相互リンクが少なく，同種のデータであるのに関連付けされているものが少ない．また，そのようなデータをまとめたポータルサイトも少ない．本研究では，この同種のデータであるが扱う地方公共団体が違うため関係付けできていないデータについて，機械的に関連性を算出し，同種のデータをまとめ，再編するシステムの実現を目指す．最終的にはオープンデータの推進や，地方公共団体の行う政策の理解の一助としたい．そこで我々は，地方公共団体の web ページから類似ページのまとまりを発見するための実験を行った．

本論文では，2 章に研究目標と目標達成に向けたアプローチ，関連研究に関して述べる．3 章ではデータのまとまりの発見を行う手順や実験対象を記述し，4 章にその実験結果を示す．5 章では今回の研究のまとめを述べる．

2. 研究目標と関連研究

本章では 2.1 に研究目標と，目標に対するアプローチについて，また 2.2 に今回用いたクラスタリング手法に関連する研究を記述する．

2.1 研究目標

本研究の目的は，各地方公共団体等が公開している基幹データの価値を高めることである．具体的には，自身の住んでいる地域の国政選挙の結果に関する情報を取得したい場合に，投票結果だけでなくその地域の投票率，全国の投票率，さらにその地域の前回選挙の結果や投票率などを関係付けて提供できる情報抽出を目標としている．なお，この目標は Linked Open Data(LOD) [1] の概念を参考にしている．本研究では，そのアプローチとして地方公共団体の web ページをクラスタリングする．次にそのクラスタを解析する手法を確立し，クラスタ間や web ページ間の関係付けを行う．最終的に同種のページだが，扱う地方公共団体が異なるため関係付けできていないページ群に対し，関係の付与を行い，またページ間の関係性を明確に示し推薦することでアクセシビリティの向上を図る．本論文

(注1)：http://statdb.nstac.go.jp/，Accessed 2013
(注2)：http://datameti.go.jp/，Accessed 2013
(注3)：http://www.data.go.jp/index.php，Accessed 2013

では，地方公共団体の web ページのクラスタリング手法や，クラスタの解析について記述する．クラスタリングの対象データは，地方公共団体の web ページ上の情報である HTML のリンクタグや，web ページのリンク構造である．

2.2 関連研究

地方公共団体の公開しているデータを対象とした先行研究はいくつかある．松村ら [2] の研究では国内の博物館情報の LOD 化を図るための LODACmuseum の活動について報告している．大城ら [3] の研究では地方議会の議事録を対象に，会議録中の意見の自動抽出を行っている．他にも web ページからの情報抽出に関して様々な先行研究が存在する．今回扱う地方公共団体の web ページに対するクラスタリングには，先行研究で行われているクラスタリング手法を参考にする．本節ではクラスタリング時に参考にした研究について記述する．

2.2.1 ストラクチャマイニングによるクラスタリング

ストラクチャマイニングとは，主に web ページのリンク構造からマイニングを行う手法である．ストラクチャマイニングにより得られた特徴量を用いてクラスタリングを行う研究を説明する．大野ら [4] [5] の研究では，Max-flow アルゴリズムを用いることで web ページの構造情報 (ストラクチャマイニング) のみで，ページの集合から小さなクラスタを生成している．さらに，生成した小さなクラスタ内のコンテンツマイニングによりクラスタの特徴量抽出を行っている．近藤ら [6] の研究では，Wikipedia のページの関連項目検索に対し，相互リンクに着目し重み付けしたグラフに対して，HITS [7] を適用する方法を提案している．これらの研究を参考に，リンク構造のみに着目したクラスタリング手法を考案し適用する．その詳細を 3.2 に記述する．

2.2.2 コンテンツマイニングによるクラスタリング

コンテンツマイニングとは，主に web ページ内の文書情報からマイニングを行う手法である．コンテンツマイニングにより得られた特徴量を用いてクラスタリングを行う研究に関して記述する．杉山ら [8] の研究では，リンク先のページの内容 (コンテンツ) を特徴ベクトルに反映させることで，クラスタリングを行い，あるページから 2〜3 リンク先の web ページにそのページの内容が集約されることを示している．阿部ら [9] の研究では，アンカーテキストを用いて既存のコンテンツベース検索システムの改善を行うことで，情報検索におけるアンカーテキストの有効性を示している．これは，web ページ内の文書でもリンク先の情報を持つことが多いアンカーテキストを用いて，文書情報と構造情報を同時に扱うことを目的としている．これらの研究を参考に，地方公共団体の任意の web ページに対する，リンク先との類似性を調べ，リンクとページ内文書の特徴の関係性抽出を行う．その詳細を 3.3 に記述する．

3. 実験の対象と手法

本章では，今回行った実験について記述する．3.1 に対象データ，3.2 にストラクチャマイニングによるページ分類実験，3.3 にコンテンツマイニングによる文書情報とリンクの関係性抽出実験について記述する．

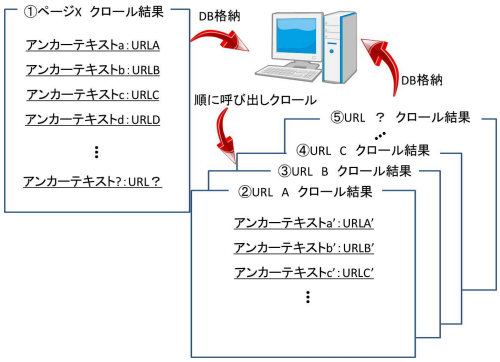

図 1 クローラーの概要

表 1 階層ごとの取得リンク数 (単位：件)

	階層 1 まで	階層 2 まで	階層 3 まで
札幌市	169	14,108	211,330
仙台市	170	7,497	110,637
さいたま市	148	16,019	388,192
横浜市	97	4,519	74,110
千葉市	115	5,510	75,342
川崎市	106	6,355	112,575
相模原市	126	4,948	84,273
新潟市	168	10,376	138,903
静岡市	148	7,025	33,922
浜松市	68	3,618	62,453
名古屋市	105	5,008	77,910
京都市	202	9,033	94,675
大阪市	110	5,834	128,498
堺市	136	8,674	96,394
神戸市	26	1,435	32,215
岡山市	208	8,718	81,541
広島市	168	9,644	149,867
北九州市	129	11,950	75,357
福岡市	173	27,525	180,902
熊本市	109	5,158	216,979

3.1 対象データ

クローラーの概要を図 1 に示す．全ての地方公共団体を対象にトップページから同一ドメイン内の全ページを対象に，リンクタグ中のアンカーテキストとそのリンク先 URL を取得するためクローラーを作成する．今回はリンク先の URL が PDF，Microsoft® Word，Excel などのファイルは解析に利用しないためクロール時に除外する．また，今回の実験はクラスタリングや解析が可能であるか確かめるためのものなので，データに関して政令指定都市 (以後，政令市)20 市に絞り，各トップページから 3 階層分のデータ取得を行った．幅優先で，取得ページにはページ番号を振っている．異なるリンクから，同一 URL を取得した際は異なるページ番号が振られデータベースに格納される．一度クロールした URL 内のデータは再取得の必要が無いため，クロール済みの URL は除外される．階層ごとのリンク数，ページ数を表 1，2 にまとめる．

3.2 ストラクチャマイニングによるページ分類実験

本実験では，ストラクチャマイニングによるページの分類実

表 2　階層ごとの取得ページ数 (単位：件)

	階層 1 まで	階層 2 まで	階層 3 まで
札幌市	170	3,525	29,073
仙台市	171	3,075	17,680
さいたま市	149	5,058	41,560
横浜市	98	3,013	27,758
千葉市	116	2,124	19,435
川崎市	107	2,519	18,136
相模原市	127	3,531	19,006
新潟市	169	3,010	25,917
静岡市	149	1,462	8,463
浜松市	69	1,279	9,215
名古屋市	106	2,281	16,794
京都市	203	2,784	18,787
大阪市	111	3,979	42,574
堺市	137	2,306	11,305
神戸市	27	640	7,525
岡山市	209	2,368	13,399
広島市	169	5,310	30,747
北九州市	130	1,907	14,852
福岡市	174	4,573	26,878
熊本市	110	2,954	25,005

験について記述する (以下，実験 1)．トップページから直接リンクが貼られているページを対象に各政令市ごとにリンク解析によるクラスタリングを行う．リンク解析によるクラスタリングでは異なるドメインのページをまとめることができないため，各政令市ごとに行ったクラスタリングの結果をトピックごとに分類する．クラスタリング，分類の詳細をそれぞれ 3. 2. 1，3. 2. 2 に記述する．

3. 2. 1　ストラクチャマイニングによるクラスタリング

ストラクチャマイニングによるクラスタリングの概要について記述する．クラスタリングには，大野ら [4] [5] の研究で用いた Max-flow アルゴリズムやリンク解析でよく用いられる PageRank [10] のスコアを参考にする．リンクを有向辺とし，政令市の web 構造の内，図 2 の構造に着目し，1 階層目の URL のリンク先ページ，1 階層目の URL で 3 階層目に出現するもののリンク元ページの一致率を特徴にクラスタリングを行う．この図 2 をグラフとして表したものが図 3 である．この図 3 で，ページ A，B のリンク元，リンク先ページに着目したものが図 4 である．A と B は一致率が高いため同じグループとすることができる．このリンク元，リンク先のページの比較を全ての 1 階層目のページに対して行い，グループの作成を行う．比較の際，PageRank の極端に高いページは，どのページとも関連があるためクラスタリング時には除外した．このグループの内，同一ページの含まれるグループの結合を行い，結合しきれなくなったものをクラスタとする． また，リンク元，リンク先の一致率のスコアには，*Jaccard* 係数を用いた．*Jaccard* 係数は 2 つの集合や 2 つのベクトルの共起度合いを表す数値である．今回の実験では，*Jaccard* 係数の要素にリンク先，リンク元 URL 群をそれぞれ用いる．例えば集合 S_1，S_2 の *Jaccard*

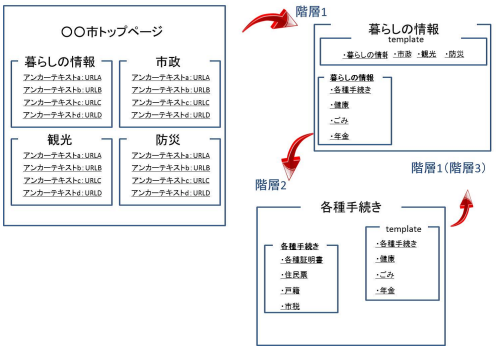

図 2　注目する構造

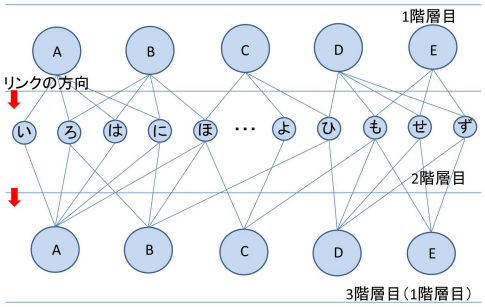

図 3　実験 1 の概要

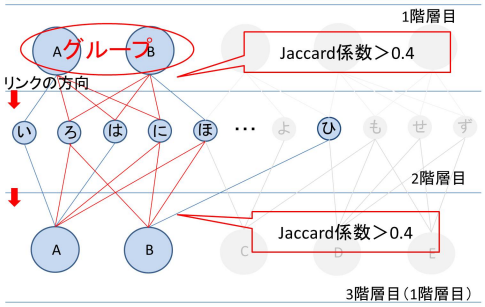

図 4　実験 1 の概要

係数を調べたいとき，S_1，S_2 内 URL の集合をそれぞれ N_A，N_B とすると *Jaccard* 係数 J は式 (1) で表される．

$$J = \frac{N_A \wedge N_B}{N_A \vee N_B} \tag{1}$$

今回は *Jaccard* 係数の閾値を 0.4 以上とした．

3. 2. 2　クラスタの分類

各政令市ごとにクラスタリングを行うことで同種のページ集合を作成するが，リンク構造のみでページ分けるため各政令市の地域サイトの構成に大きく依存した特徴を持つクラスタが生成される．異なる地域サイトの同種クラスタをまとめる際，これらを画一的な指標でまとめる必要がある．このことを考慮し，クラスタの分類を行う．分類器にはベイズ分類器を用いる．またベイズ分類器の教師データとして LDA(Latent Dirichlet Allocation) [11] により作成したトピックを用いる．政令市 20 市の 1 階層目のページの全文書に対し，LDA を用いて 50topics，5words のトピック群の作成を行った．これによりそれぞれの地域サイトの構成に依らない指標を作成する．このトピック群を元にクラスタの分類を行う．

3.3 コンテンツマイニングによるリンク間関係性抽出実験

本実験では，リンク-被リンクの関係にあるページ間の文書を比較し，類似度を算出することで，リンクとページ内文書の特徴の関係性抽出を行う．本実験の目的は，リンク間のページの特徴の類似性を示すことで，3.2 で記述した実験 1 を全てのリンク元，リンク先の URL を取得せずとも，ある程度分類できる様にすることである．一般的に同一カテゴリ内のページ間の類似度は，他のページ間の類似度と比べ高く，またリンクの繋がるページ間には関係性があると考えられる．そこで，京都市の地域サイトの 3.1 で取得したデータを対象に，リンク-被リンクの関係にあるページ間のアンカーテキストの比較を行う．コンテンツマイニングによるリンク間関係性抽出実験の概要について説明する (以下，実験 2)．リンク間の比較にはコサイン類似度を用いる．コサイン類似度の要素にはページ内単語の $tfidf$ 値から作成したベクトルを用いる．そのベクトル同士の成す角度の小ささで類似度を判定し，コサイン類似度が 1 に近いほど類似している．コサイン類似度 $cos(\vec{a},\vec{b})$ は，式 (2) で表される．

$$cos(\vec{a},\vec{b}) = \frac{\vec{a}\cdot\vec{b}}{|\vec{a}|\cdot|\vec{b}|} \tag{2}$$

なお，ベクトル生成に用いた $tfidf$ 値の算出の際，ページ内アンカーテキスト群を文書 d とし，総文書数 (総ページ数) を D，単語 i の文書 d_j における出現回数を n_{ij}，文書 d_j 内の総単語数を N_j，単語 i を含むドキュメント数を d_i とする．このドメイン内における tf 値，idf 値は式 (3)，(4) で表される．

$$tf = \frac{n_{ij}}{N_j} \tag{3}$$

$$idf = log\frac{D}{d_j} \tag{4}$$

$tfidf$ 値は，単語の出現頻度である tf 値と逆文書頻度 idf 値の積として式 (5) で表される．

$$tfidf = tf \cdot idf \tag{5}$$

リンク-被リンク間は 1 対多の関係にあるため，リンク元ページの特徴量としてリンク先ページとの類似度の平均を用いる．図 5 に平均コサイン類似度算出例を示す．ページ A とそのリンク先ページ $(a_1, a_2, a_3, ..., a_n)$ をそれぞれ比較する際，式 (2) のコサイン類似度を用いる．ページ A の特徴量 F は，式 (6) で表される．

$$F = \frac{1}{n}\sum_{i=1}^{n} cos(\vec{A},\vec{a_i}) \tag{6}$$

$A, a_1, a_2, a_3, ..., a_n$ に出現する名詞とその $tfidf$ 値でベクトル空間を作成し，$cos(\vec{A},\vec{a_1}), cos(\vec{A},\vec{a_2}), cos(\vec{A},\vec{a_3}),..., cos(\vec{A},\vec{a_n})$ の平均をページ A の特徴量 F とする．このページ A の特徴量 F を平均コサイン類似度と呼ぶ．

4. 実験結果と考察

本章では，4.1 で実験 1 の結果，4.2 で実験 2 の結果，4.3 に考察をそれぞれ記述する．

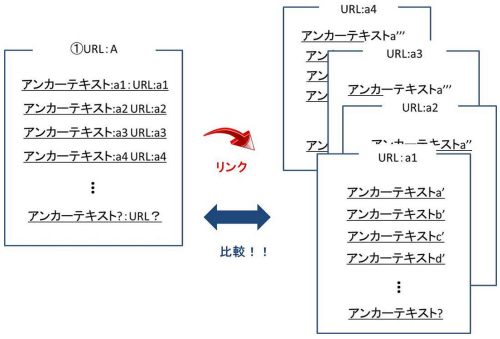

図 5　平均コサイン類似度算出例

表 3　静岡市 1 階層目のクラスタリング結果

カテゴリ名	クラスタ内 URL 数	総 URL 数	precision	recall
都市計画・建設	8	8	1.00	1.00
観光・イベント	5	5	1.00	1.00
行財政	2	5	1.00	0.40
産業・労働	6	6	1.00	1.00
消防・防災	7	9	1.00	0.78
地域づくり	7	8	1.00	0.88
静岡市長の部屋	3	3	1.00	1.00
新着情報	8	8	1.00	1.00
行財政	3	5	1.00	0.60
医療・保健衛生	9	11	1.00	0.82
くらし	6	10	1.00	0.60
教育・文化	6	8	1.00	0.75
環境	9	11	1.00	0.82

表 4　京都市 1 階層目のクラスタリング結果

カテゴリ名	クラスタ内 URL 数	総 URL 数	precision	recall
まちづくり	5	11	1.00	0.45
健康・福祉・教育	10	16	1.00	0.63
観光・文化・産業	13	14	1.00	0.93
暮らしの情報	18	19	1.00	0.95
交通	2	2	1.00	1.00
市政情報	19	22	1.00	0.86
上京区	2	3	1.00	0.67
カレンダー	15	34	1.00	0.44

4.1 実験 1 の結果

政令市 20 市に対し，リンク構造に着目したクラスタリングを行なった．クラスタリングの結果，静岡市，京都市，岡山市，福岡市で同種のページをまとめることができた．地域サイトにおける区分に対するクラスタの適合率，再現率を表 3，4，5，6 にまとめる．

京都市を取り上げて記述する．“まちづくり”，“健康・福祉・教育”，“観光・文化・産業”，“暮らしの情報”，“交通”，“市政情報”，“上京区”，“カレンダー” のページ集合がまとまった 8 つの小さなクラスタに分けることができた．

次に，クラスタリングに成功した 4 市の 1 階層目のページに対して同種のページをまとめる．そのために，まず政令市 20 市 1 階層目のページ内全文書に対して LDA を用いてトピック抽

表 5　岡山市 1 階層目のクラスタリング結果

カテゴリ名	クラスタ内 URL 数	総 URL 数	precision	recall
市政情報	15	17	0.93	0.82
注目キーワード	8	12	1.00	0.67
観光・文化・イベント	8	12	1.00	0.67
くらしのイベント	8	12	1.00	0.67
施設案内	18	19	1.00	0.95
事業者情報	5	10	1.00	0.50
くらしの情報	16	20	1.00	0.80

表 6　福岡市 1 階層目のクラスタリング結果

カテゴリ名	クラスタ内 URL 数	総 URL 数	precision	recall
くらし・手続き・環境	18	18	1.00	1.00
経済・産業・ビジネス	8	13	1.00	0.62
観光・イベント・魅力	8	18	1.00	0.44
関連メールマガジン	13	13	1.00	1.00
市政情報・市民参加	18	17	0.94	1.00

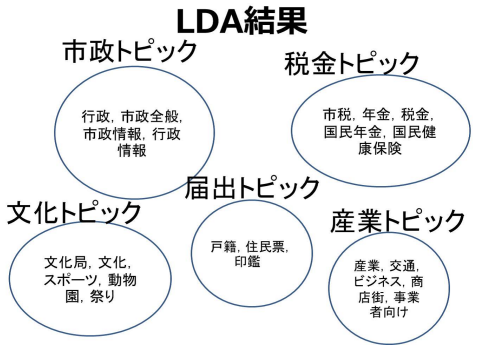

図 6　LDA の結果

出を行う．抽出できた主なトピックを図 6 に示す．抽出したトピックを確認したところ，主なトピックとして市政，税金，文化，届け出関係，産業などに関する単語がまとまっているものが確認できた．このトピックを教師データとしてベイズ分類器でクラスタの分類を行った．結果としては市政，税金トピックに，それぞれの政令市の市政情報や税金カテゴリのクラスタが割り振られていた．

4.2　実験 2 の結果

京都市の地域サイトを対象に，実験 2 を行った結果を記述する．階層別の平均コサイン類似度の統計量を表 7 に示す．平均コサイン類似度の平均は，階層が深くなるほどわずかに減少している．しかし，分散は階層が深くなるほどわずかに大きくなっている．また，表 8 に京都市の地域サイトにおけるコサイン類似度と階層数の相関係数を示す．相関係数 r(対象ページ数 $n = 11,030$) は，-0.03 となった．これより非常に弱い負の相関が認められる．

次に 4.1 より得られた京都市のクラスタごとに平均コサイン類似度の平均を算出した．結果を表 9 に示す．これよりリンク解析により得られたカテゴリごとのクラスタは比較的リンク間

表 7　京都市の階層別平均コサイン類似度の統計量

	0 層→ 1 層	1 層→ 2 層	2 層→ 3 層
平均	0.117	0.062	0.055
中央値	0.117	0.050	0.040
最頻値	—	0.040	0.041
分散	—	0.002	0.003

表 8　京都市の地域サイトにおける平均コサイン類似度と階層数の相関係数 (対象ページ数 n=11,030)

	階層数	平均コサイン類似度
階層数	1	
平均コサイン類似度	-0.031	1

表 9　クラスタ別平均コサイン類似度の平均

カテゴリ名	平均コサイン類似度の平均	URL 数
まちづくり	0.127	5
健康・福祉・教育	0.114	10
観光・文化・産業	0.109	13
暮らしの情報	0.094	18
交通	0.087	2
市政情報	0.081	19
上京区	0.069	2
カレンダー	0.034	15

のページ文書の類似度が高く，カレンダーのページなどのテンプレートページのクラスタは比較的リンク間のページ文書が低いことが確認できる．

4.3　考　　察

実験 1 のクラスタリンクでは，静岡市，京都市，岡山市，福岡市の地域サイト 1 階層目の URL に関して，高い適合率でクラスタを生成することができた．今回扱った手法で，精度の高いクラスタリングが行えたのは，それぞれの地域サイトのリンク構成に依るところが大きいと考えられる．他にも様々なリンク構造について注目して再現率の向上を図る必要がある．分類に関しても，市政情報カテゴリや税金のカテゴリが集まるクラスタを同種クラスタとして分類することができた．地方公共団体の地域サイトは他にも多彩なトピックを持っており，現状 2 つのトピックのページ群の関係付けしかできていないが，今後分類方法の見直しや LDA によるトピックの抽出の精度を向上させることで，他のトピックにも対応させていく必要がある．

実験 2 のリンク間関係性抽出実験では，平均コサイン類似度の平均は階層が深くなるに従い，わずかに減少している．分散は階層が深くなるに従い，わずかに増加しているが無視できる程度である．また，表 8 より相関係数 r =-0.03 であるから，平均コサイン類似度と階層数の間には非常に弱い相関が見られる．しかし，この数値も無視できる程度である．また表 9 から京都市の地域サイトに関して，何らかのカテゴリのページの方がテンプレートページより比較的リンク間の文書が似ていることがわかる．そのため，類似度を上げるためには何らかのカテゴリのページのリンクとして存在するテンプレートページの除去を行う必要があることがわかった．今回の実験ではリンク間の文

書の類似度を示す程度に留まったが，今後はページのトピックの推定やリンク間でのトピックの関係性 (上下関係，同位関係等) を示すことを目標として行きたい．

5. おわりに

本稿では，情報抽出分野で与えられた検索クエリに対しての情報推薦手法を，サンプルデータではなく現存する特定ドメインに対して適応した．また，全ての地方公共団体の，全てのページを対象にすることはできなかったが，政令指定都市 20 市の内，1 階層目の URL に対して，4 市の地域サイトに関してそれぞれのサイトの構成による形でクラスタリングする事ができた．今後，このような膨大なデータを処理する上で，実験 2 による仮説の実証が不可欠となる．そのため，今回は京都市の 3 階層分のデータを用いて，1 階層目の URL に対して解析するのみにとどまったが，他の地方公共団体や 3 階層以上の URL に対しても解析を行い，リンク間のトピックの関係性を示す必要がある．また，仮説が実証できなかった際に，このような膨大なデータの処理をする何らかの手法も合わせて検討する必要がある．最後に，実験を行っていく中で行政機関のデータの不統一性を改めて実感した．取得 URL の中に web 上で検索をしても見つからないページや，同一ドメイン内で文書の文字コードが異なっているものも確認できた．ページが見つけられないものは，ソースコードの相対パスなどの設定が正しくない，リンクが切れているなど原因は様々であった．文字コードも多言語を扱う地方公共団体には仕方のない問題ではある．これらはクローラーの改良にて対応できる可能性がある．

謝　辞

本研究の一部は，首都大学東京傾斜的研究費「関心に誘発されるデータマイニングの研究」によって行われたものである．

文　献

[1] Bizer.C., Heath.T. and Berners-Lee.T.,「Linked Data The Story So far」, International Journal on Semantic Web and Information Systems(IJSWIS), Vol. 5, No. 3, pp. 1-22, 2009.

[2] 松村 冬子，嘉村 哲郎，加藤 文彦，小林 厳生，高橋 徹，上田 洋，大向 一輝，武田 英明「LODAC Museum: Linked Open Data による博物館情報の統合と活用」，The 26th Annual Conference of the Japanese Society for Artificial Intelligence, 2012，3C2-OS-13b-9.

[3] 大城 卓，渡邉 祐斗，渋木 英潔，木村 泰知，森 辰則,「地方政治情報システムのための地方議会会議録への注釈付けタグセットの提案」，言語処理学会，第 18 回年次大会，発表論文集 (2012 年 3 月).

[4] 大野 成義，太田 学，片山 薫，石川 博,「特徴間の類似性を考慮した特徴量集約手法の検討」，電子情報通信学会 第 18 回データ工学ワークショップ論文集 (DEWS2007)，論文集，2007.

[5] 大野 成義，渡辺　匡，片山 薫，石川 博，太田 学,「Max Flow アルゴリズムを用いた Web ページのクラスタリング方法とその評価」，情報処理学会論文誌，データベース 47(SIG 4(TOD 29))，65-75，2006-03-15.

[6] 近藤 直樹，渡辺 陽平，横田 治夫,「Wikipedia のセクションを考慮したリンク解析による関連項目検索手法の提案」，全国大会講演論文集 第 72 回平成 22 年 (5)，“5-183”-“5-184”，2010-03-08.

[7] J. M. Kleinberg,「Authoritative sources in a hyperlinked environment」, J. ACM, pp．604632, 1999.

[8] 杉山一成，波多野賢治，吉川正俊，植村俊亮,「リンク先ページの内容を反映させた Web ページの特徴ベクトル」，電子情報通信学会第 13 回データ工学ワークショップ (DEWS2002)，A1-3，Mar.

[9] 阿部匡史，豊田正史，喜連川優,「アンカーテキストとリンク構造解析を用いた Web 情報検索の改善」，電子情報通信学会第 14 回データ工学ワークショップ (DEWS2003)，論文集，2003.

[10] S. Brin and L. Page,「The anatomy of a large-scale hypertextual (web) search engine」, Computer Network and ISDN Systems, 30(1-7)：107-117, 1998.

[11] D. M. Blei, A. Y．Ng, and M. I. Jordan, 「Latent Dirichlet Allocation」, Neural Information Processing Systems 14, 2001.