Wikipediaを用いた観光オブジェクトの属性抽出に基づく対応Webページの特定手法

峯　　祥平†　　北山　大輔††

† 工学院大学大学院工学研究科情報学専攻　〒163-8677 東京都新宿区西新宿1丁目24番地2号
†† 工学院大学情報学部コンピュータ科学科　〒163-8677 東京都新宿区西新宿1丁目24番地2号
E-mail: †em14013@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　一般に，ユーザが観光の計画を立てる際，ガイドブックやWebから情報を収集するが，観光に関する情報は多数存在し情報同士の関連付けがされてないことが多いため，情報収集のために手動で閲覧し取捨選択しなければならない．こういった手間の削減のために，ユーザの定型的な行動を他オブジェクトに対しても適応することで，手軽に情報を収集できるスマートスクラップブックシステムを提案した．本稿では，ユーザが，あるWebサイトにおける概要やアクセス情報などのWebページの一部を保存した場合，他オブジェクトのWebサイトにおいて，保存したWebページの一部に対応するWebページを特定する手法を提案する．具体的には，種類に応じた対応情報の特定を行う．ユーザが手動で保存した内容と，その観光オブジェクトに対応するWikipediaページの概要，歴史などの各項目との類似度を測る．最も類似する項目名を保存内容の属性として抽出する．次に，他オブジェクトのWikipediaページ内にある，該当属性の項目内容と類似するWebページを抽出する．

キーワード　観光情報，属性抽出，対応関係抽出，Wikipedia

1. はじめに

近年，ユーザが観光の計画を立てる際，目的とする観光オブジェクトに関する情報をWebやガイドブックを用いて収集することが一般的となっている．観光オブジェクトに関する総合的な情報が載っている公式サイト，過去にそこを訪れた観光者による旅行記ブログ，そして，評価を載せたレビューサイトなど様々な情報がWeb上に存在する．そのため，ガイドブックによる情報のみでは不足する時，こういった情報を補足情報として収集することが可能である．しかし，こういった情報は混在しており，さらには情報同士の関連付けがされてないため，目的の情報を収集するためには手動で閲覧し取捨選択しなければならない．そこで我々は，Webで詳細を調べるなどの行動には特定のパターンがあると考え，そのパターンを他のオブジェクトに対しても適応することで，手軽に観光情報を収集できるスマートスクラップブックシステムを提案した[1]．

本稿では，図1のように，ユーザが保存した内容の属性を判定し，その属性に対応した別のオブジェクトにおける内容を発見するための対応Webページ特定手法を提案する．本手法では，まずユーザが，あるWebサイトにおける概要やアクセス情報などのWebページの一部を保存する．次に，保存された内容とWikipediaの概要や歴史といった各項目ごとの類似度を計算し，最大となる項目を，その保存した内容における属性と定める．そして，ユーザが他オブジェクトのスクラップブックを作成した時に，システムは他オブジェクトにおけるWikipediaページ上から属性と一致する項目を発見，内容を抽出する．さらに，Web検索結果のサイト内のページ集合に対して総当りで，抽出されたWikipediaの内容との類似度を計算する．また，同様に，総当りでユーザが保存した内容との類似度も計算することで，拝観料などのWikipediaには存在しない項目の情報や，アクセスのような「駅」「乗車」といった，ほとんどの観光オブジェクトに共通する名詞が含まれるような内容にも対応できるようにした．これらの類似度を乗算しスコアリングし，スコアが最も高くなるページを自動的に保存する．

以下，本論文の構成を示す．まず，2節では本研究の関連研究について説明する．3節では対応Webページ特定手法について説明する．4節では属性の判定，5節では対応Webページ特定手法の，それぞれの評価実験について説明する．

2. 関連研究

2.1 文書の属性や関係に関する研究

文書解析から属性や関係性を抽出し，適応する研究は様々ある．小谷ら[2]は，複数のWebサイト間における共通属性を抽出し，その共通属性の各属性に該当するWebページを抽出する手法を提案している．ユーザが複数のサイトを指定する必要がある点や，手法において多量のサイトからsiteFrequencyを利用する点が特徴だが，我々の研究ではユーザが観光のための情報収集を前提としているため，こういったケースには十分に対応できないと考えられる．

佃ら[3]は，オブジェクトと属性値の関係の認知度を推定するための手法を提案している．2つのオブジェクトの認知度と関連度といった指標を用いて，意外な共通点を発見する．我々の研究ではユーザとシステムによって観光オブジェクトの情報が複数保存された時，位置関係が分かる地図も同時に保存することを今後の課題としている．複数のオブジェクトを横断した情報の発見という点で類似しており，さらに本研究のシステム

により，観光地の意外な関係性といった情報を自動的に追加する提案はオブジェクトへの興味を更に刺激する点でとても有用であると考えられる．

田中ら [4] は，ユーザに対してニュース記事の理解を支援するために，ニュース記事から人物や地域などのエンティティを抽出し，エンティティ間の関係を示す情報を取得することで，記事のための背景知識を補完するための手法を提案している．文書解析によって，ユーザの関心のある新たな情報を発見し取得する点では類似するが，我々の研究で扱うものは観光情報であり，さらに，他オブジェクトへの対応する内容の保存という点で異なる．

加藤ら [5] は，異なるドメインにおいて例示検索を行う方法を提案している．表現方法は異なるが，対応する情報を提示するという点では類似しているが，我々の研究では入力キーワードではなく，その検索結果に対するユーザの検索操作パターンから関係性を抽出し，提示する情報は文章であるため手法が異なる．

2.2 観光情報に関する研究

また，ユーザが観光情報を取得する際の手助けとなる研究もいくつか存在する．三笠ら [6] は，旅行記の観光トピックごとに文章分類を行い，動的に概要文章を生成する手法を提案している．これにより，閲覧者は興味に応じた要約文章を見ることができ，不必要だと事前の判断が可能になるためサイトの閲覧時間を削減できる．観光情報の文書分類をする点では類似するが，あくまでも我々の研究では他オブジェクトへの反映がメインであるため，本手法ではうまくいかないことが予想される．また，我々の研究では保存する内容の文章量は考慮されないため，スクラップブックの閲覧性に関しては本研究による要約された文章のようなものを提案する必要がある．

石野ら [7] らは，旅行記が記述されたブログエントリから自動的にリンクを収集し分類，それにより低コストでの観光情報リンク集の構築するための手法を提案している．これにより，歴史やニュースなどの幅広い情報へのリンクなどを自動的に補填できる．観光旅行を目的としている点や情報の分類という点で類似している．本研究では一覧性のあるリンク集を構築しているが，我々の研究ではユーザの求めている情報のみを集約するという点で異なる．

3. 対応 Web ページ特定手法

本節では，ユーザがあるオブジェクトに関して手動で Web 検索を行い，結果内のあるページの一部をスクラップブックに追加した場合にそれを他のオブジェクトに適応させる手法を説明する．ユーザが観光オブジェクト A について Web 検索を行い，ある Web サイトにおける概要やアクセスなどの Web ページの一部のテキスト T_X を保存したとする．この時，ユーザは他オブジェクトの Web サイトにおいても同種の情報を求めると考えた．さらに，他オブジェクトにおいて，同種の情報は以下の特徴を持つと考えた．

- コンテンツが類似している

これは，アクセス情報や料金情報のように，オブジェクトに依存しない移動手段や，その情報を表現する語句が類似する場合に見られる特徴である．

- 属性が一致している

これは，たとえば，オブジェクトの歴史を保存した場合，同様に歴史の情報が対応するということである．この時，コンテンツそのものが類似するとは限らない．

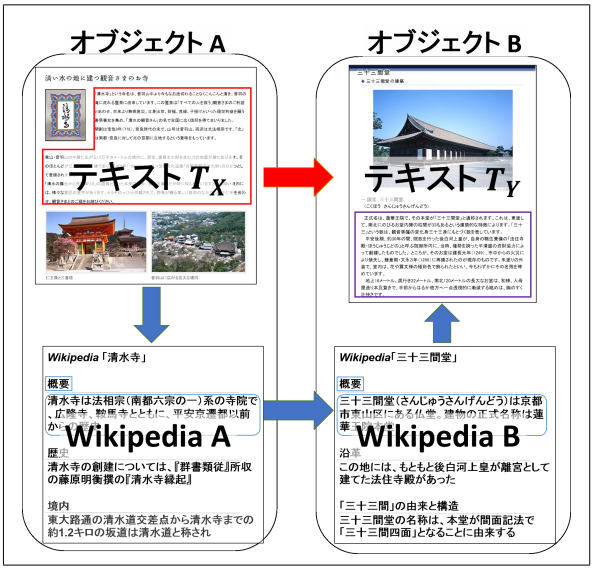


図 1 対応 Web ページ特定手法の概念図

この 2 つの特徴のうち，属性が得られるならば，属性の一致による判断を使うべきだと考えられるが，一般に属性は明示的に与えられることはない．そのため，本手法では，この 2 つの指標を組み合わせて用いる．

3.1 コンテンツの類似

コンテンツの類似について説明する．まず，ユーザがスクラップブックに保存したオブジェクト A に関するテキスト T_X から，形態素解析エンジン Mecab [8] によって名詞を抽出し，種類と個数から特徴ベクトル X を生成する．また，$X = \{x_1, x_2, ..., x_n\}$ である．

$$x_i = \frac{count(T_{Xi}, T_X)}{\sum_{T_{Xj} \in T_X} count(T_{Xj}, T_X)} \quad (1)$$

$count$ は，T_X における単語 T_{Xi} の個数を求める関数である．次に，ユーザが他オブジェクト B に関してスクラップブックを作成した場合，システムはオブジェクト B に関する検索結果の上位 K 位までのページと，そのページから L リンク先までをページ集合 P として取得する．取得したページ全てに対しても特徴ベクトルを生成する．そして，T_X から P のそれぞれに対して総当りでコサイン類似度 [9] を求めた結果，上位にあるものを類似したコンテンツとする．

以下の数式によりコンテンツの類似度を算出する．

$$cos(X, Y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2} \sqrt{\sum_{i=1}^{n} {y_i}^2}} \quad (2)$$

$cos(X, Y)$ は，観光オブジェクト A において，ユーザが保存したテキスト T_X における名詞ベクトル X と，観光オブジェクト B において保存される候補テキスト T_Y における名詞ベクト

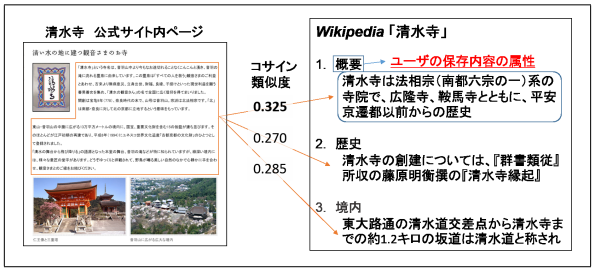


図 2　清水寺における属性判定例

図 3　清水寺における属性を利用した対応情報の保存

ル Y の類似度を算出するための式である．ベクトル X の要素は x，ベクトル Y の要素は y である．また，テキスト T_X とテキスト T_Y の全名詞集合の要素数は n である．

3.2　属性の判定

属性の判定について説明する．ここでは，オブジェクトに対応する Wikipedia を用いて属性の判定を行う．Wikipedia 記事間の項目の対応関係を利用することで，内容の類似度だけでなく，属性の一致度という指標から同種の情報を発見できると考えられるためである．まず，3.1 節と同様に，ユーザがスクラップブックに保存したオブジェクト A に関するテキスト T_X から名詞を抽出し，特徴ベクトル X を生成する．この時，Wikipedia A の各項目ごとの特徴ベクトルを生成し，これらに対して特徴ベクトル X とのコサイン類似度を求める．類似度が最大となった項目の特徴ベクトルを W とし，その内容が書かれた項目名 N を T_X における属性と定める．式 2 を用いて，$cos(X, W)$ を求めることで，W を抽出する．あるユーザが清水寺に関して内容を保存し，次に三十三間堂のスクラップブックを作成した場合を例に説明する．まず，ユーザによって保存された図 2 のような内容と Wikipedia の各項目との類似度を算出し，最大となった内容をもつ項目名「概要」が属性となる．

3.3　属性の一致度

ユーザが他オブジェクト B に関してスクラップブックを作成した場合，ユーザが保存した T_X の属性 N がユーザの興味であると考えられるため，オブジェクト B に対応する Wikipedia B から Wikipedia A の項目名 N と同名の項目名 N' を抽出する．また，システムはオブジェクト B に関する検索結果の上位 K 位までのサイトのページと，そのページから L リンク先までをページ集合 P として取得する．サイト単位でページ集合を取得する理由は，最上位のサイトのみでは情報に偏りができると考えられるためである．そして，項目名 N' のテキスト W' から P に対して総当りでコサイン類似度を求めた結果，上位にあるものは一致した属性をもつ内容である．式 2 を用いて，$cos(W', Y)$ を求めることで，T_Y を抽出する．この時の

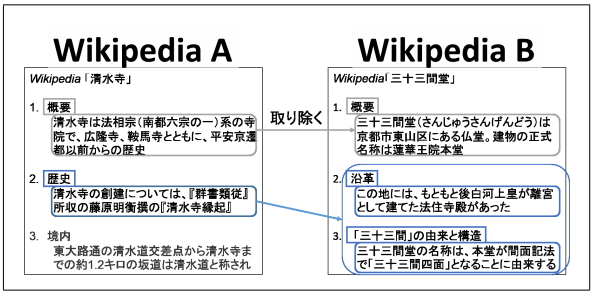
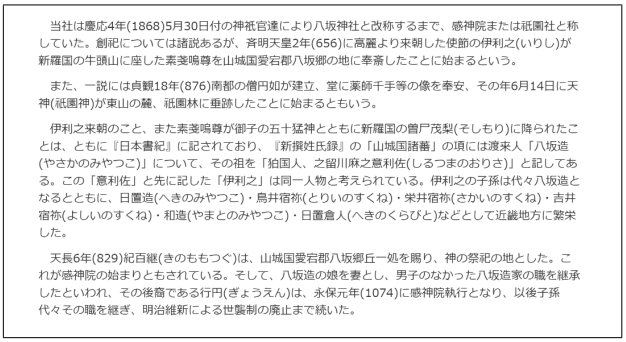


図 4　項目が存在しない場合の対応関係


当社は慶応4年(1868)5月30日付の神祇官達により八坂神社と改称するまで、感神院または祇園社と称していた。創祀については諸説あるが、斉明天皇2年(656)に高麗より来朝した使節の伊利之(いりし)が新羅国の牛頭山に座した素戔嗚尊を山城国愛宕郡八坂郷の地に奉斎したことに始まるという。

また、一説には貞観18年(876)南都の僧円如が建立、堂に薬師千手等の像を奉安、その年6月14日に天神(祇園神)が東山の麓、祇園林に垂跡したことに始まるともいう。

伊利之来朝のこと、また素戔嗚尊が御子の五十猛神とともに新羅国の曽尸茂梨(そしもり)に降られたことは、ともに『日本書紀』に記されており、『新撰姓氏録』の「山城国諸蕃」の項には渡来人「八坂造(やさかのみやつこ)」について、その祖を「狛国人、之留川麻之意利佐(しるつまのおりさ)」と記してある。この「意利佐」と先に記した「伊利之」は同一人物と考えられている。伊利之の子孫は代々八坂造となるとともに、日置造(へきのみやつこ)・鳥井宿祢(とりいのすくね)・栄井宿祢(さかいのすくね)・吉井宿祢(よしいのすくね)・和造(やまとのみやつこ)・日置倉人(へきのくらびと)などとして近畿地方に繁栄した。

天長6年(829)紀百継(きのももつぐ)は、山城国愛宕郡八坂郷丘一処を賜り、神の祭祀の地とした。これが感神院の始まりともされている。そして、八坂造の娘を妻とし、男子のなかった八坂造家の職を継承したといわれ、その後裔である行円(ぎょうえん)は、永保元年(1074)に感神院執行となり、以後子孫代々その職を継ぎ、明治維新による廃止まで続いた。

図 5　「八坂神社の歴史｜八坂神社」内の一部

$cos(W', Y)$ の値を属性の一致度と呼ぶ.「概要」と同名の項目名を適応オブジェクトである三十三間堂の Wikipedia より発見する場合を例に説明する．そして図 3 のように，三十三間堂の検索結果上位 K 件のページ集合 P に対して総当りで，Wikipedia の項目「概要」の内容との類似度を各ページごとに算出し保持する．

なお，閾値を 0.2 と定め，全項目とのコサイン類似度の値が閾値未満しか存在しない場合は，一致する属性がないと考えられるため内容の類似度のみを対応度に用いる．また，2 つのオブジェクトにおける Wikipedia 記事間の項目の対応関係を見つける際，同名の項目が存在しない場合は，図 4 のように全項目を 1 つの項目と見なす．その際，Wikipedia B において，Wikipedia A に存在する項目名と一致するものは取り除く．

3.4　対応ページの特定

最後に，3.1 節で求めた「コンテンツの類似度」と 3.3 節で求めた「属性の一致度」を平均し，最も対応度が高くなったページがユーザの保存した内容に対応する情報だと考えられるため，スクラップブックに保存する．

以下の数式により同種の情報の対応度を算出する．

$$Score(X, Y) = \frac{cos(W', Y) + cos(X, Y)}{2} \qquad (3)$$

$Score(X, Y)$ は，ユーザが保存した観光オブジェクト A におけるテキスト T_X と，システムによって保存される観光オブジェクト B における候補テキスト T_Y の対応度を求める式である．

4.　評価実験：属性の判定

属性の判定に関して評価するために，被験者によって，以下の手順で正解を作成した．

（1） 八坂神社に関する 3 つの保存内容を見て，それぞれの内容を表すようなキーワードやタイトルを自由記述で書く．

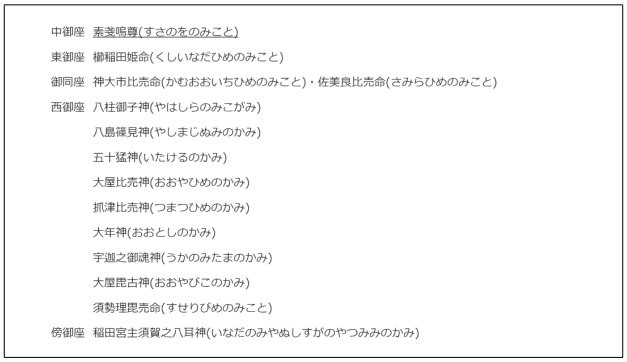


図 6 「御祭神｜八坂神社」内の一部

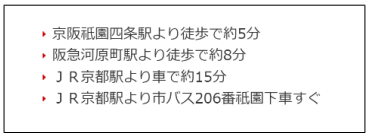


図 7 「境内マップ・アクセス｜八坂神社」内の一部

（2） Wikipedia 内の八坂神社のページ項目一覧のみが書かれた表を見て，八坂神社に関する 3 つの内容に関して (1) で回答したキーワードに該当すると考えられる項目名を書く．
被験者 10 名の回答を元に，最も回答が多いものを正解と定める．

考察の観点として，まず，ユーザが自由記述で書いたキーワードと，選択した Wikipedia の項目との比較を行うことで，属性の判定に Wikipedia を使用することの妥当性を評価する．また，属性の判定によって最上位になった属性と，ユーザの選択した Wikipedia の項目の一致により，判定手法の妥当性を確認する．

4.1 実験データ：属性の判定

実験を行うにあたり，使用したデータを説明する．観光オブジェクト「八坂神社」に関して，図 5，図 6，図 7 の 3 つの内容を使用する．それぞれ，「八坂神社の歴史 (注1)」，「御祭神 (注2)」，「境内マップ・アクセス (注3)」のページ内から抽出した．表 1 に示すように，本手法によって類似度が最大 (下線部) となった Wikipedia の項目名を属性とした．

4.2 考察：属性の判定

「八坂神社の歴史」，「御祭神」，「境内マップ・アクセス」の内容に関して，被験者が書いた自由記述のキーワードと選択した Wikipedia の項目名を表 2 に示す．自由記述では，被験者の回答は一致が見られないが，Wikipedia の項目名としては，大多数が同じ項目を選択した．そのため，Wikipedia に該当項目が存在する場合は，属性名を Wikipedia の項目から抽出することは妥当であると考えれる．また，表 1 でシステムによって得られた属性と，表 2 の被験者の選択した Wikipedia の項目名は全て一致していた．以上のことより，属性の判定手法は妥当であると考えられる．

(注1)：http://www.yasaka-jinja.or.jp/about/
(注2)：http://www.yasaka-jinja.or.jp/about/saijin.html
(注3)：http://www.yasaka-jinja.or.jp/access.html

表 1 Wikipedia 八坂神社の各項目と，保存内容とのコサイン類似度

	八坂神社の歴史	御祭神	境内マップ・アクセス
項目「社名について」	0.31	0.09	0.17
項目「概要」	0.17	0.03	0.17
項目「祭神」	0.31	0.76	0.01
項目「歴史」	0.44	0.10	0.09
項目「摂末社」	0.15	0.22	0.03
項目「主な祭事」	0.12	0	0
項目「文化財」	0.28	0.02	0.07
項目「現地情報」	0.04	0.02	0.48
項目「脚注」	0.23	0.35	0.02
項目「参考文献」	0.17	0	0
項目「関連項目」	0.02	0	0.09
項目「外部リンク」	0.04	0	0

5. 評価実験：対応ページの特定

対応 Web ページ特定手法に関して評価するために，被験者によって，正解を作成した．被験者は，八坂神社に関する 3 つの内容に対応しているものを京都タワー，伏見稲荷大社，東福寺，それぞれのページ集合から 1 つずつ選択した．また，八坂神社に関する 3 つの内容は 4 節と同様のものを使用した．被験者 10 名の回答を元に，最も回答が多いものを正解と定める．

考察の観点として，被験者が，京都タワー，伏見稲荷大社，そして，東福寺に関して，八坂神社の 3 つの内容に対応していると判断したページと，対応ページ特定手法によって算出したページを比較し，評価する．また，この 3 つの観光オブジェクトを，以降は適応オブジェクトと示す．

5.1 実験データ：対応ページの特定

実験を行うにあたり，使用したデータを説明する．実行環境は Google 検索エンジンである．データセットとして「八坂神社」，「京都タワー」，「伏見稲荷大社」，「東福寺」の 4 件の観光オブジェクトを用意した．具体的には，八坂神社に関しては，4. 節と同様の 3 つの内容を利用する．適応オブジェクトに関しては，検索クエリを観光オブジェクト名とした際の検索結果の最上位ページとその 1 リンク先のページ集合である．あらかじめ，八坂神社の内容から適応オブジェクトに対して，対応ページ特定手法を適応し，式 3 の Score を算出する．京都タワー，伏見稲荷大社，東福寺の各ページの Score は，表 3，表 4，表 5 である．これらは，各ページの内容に対して，Wikipedia 上の属性の内容と八坂神社の対応ページ，それぞれのコサイン類似度とその平均値である Score を算出した結果である．各表は本手法を使用した際の精度の降順でページ 5 件ずつとなっている．表中のページタイトルは短いテキストは実際のものを使い，長いテキストや複雑な表記のものに関しては適宜編集したものを記載している．被験者は，このページ集合から対応していると考えられるページを判断する．また，属性「複合」は，同一属性が存在しないために，共通内容を削除し，それ以外を複合したものである．

5.2 考察：対応ページの特定

ユーザが，京都タワー，伏見稲荷大社，そして，東福寺に関して，八坂神社の 3 つの内容に最も対応していると判断した

表 2　八坂神社の各内容に関する被験者の回答

保存内容	内容を表すキーワードやタイトル (回数)	選択項目名	項目別人数
八坂神社の歴史	八坂神社の歴史 (2)，歴史 (2)，神社の歴史 (1)，八坂神社 (1)，京都 (1)，八坂神社の沿革 (1)，なり立ち (1)	歴史	9 人
	概要 (1)	概要	1 人
御祭神	神 (2)，神様の名前 (1)，神様の一覧 (1)，八坂神社の御神体 (1)，日本神話 (1)，和神 (1)，御座 (1)，祭神 (1)	祭神	9 人
	八坂神社の役職 (1)	摂末社	1 人
境内マップ・アクセス	アクセス (3)，神社の場所 (1)，アクセスマップ (1)，八坂神社の交通案内 (1)，所在地 (1)，清水寺 (1) 八坂神社へのアクセス (1)	現地情報	9 人
	アクセス (1)	該当なし	1 人

表 3　京都タワーのページ集合における，各コサイン類似度と Score

ページタイトル	属性「複合」	八坂神社の歴史	Score
タワーについて	0.75	0.05	0.40
セット入場券-販売	0.66	0.11	0.39
大浴場 営業時間	0.62	0.04	0.33
お知らせ一覧	0.61	0.04	0.32
バレンタイン	0.62	0.02	0.32

ページタイトル	属性「複合」	御祭神	Score
タワーについて	0.75	0.01	0.38
セット入場券-販売	0.66	0.00	0.33
タワーリンク	0.63	0.00	0.32
おたべなさいだー	0.63	0.00	0.32
交通アクセス	0.63	0.00	0.32

ページタイトル	属性「複合」	境内マップ・アクセス	Score
タワーについて	0.75	0.19	0.47
交通アクセス	0.63	0.25	0.44
セット入場券-販売	0.66	0.16	0.41
おたべなさいだー	0.63	0.17	0.40
イベント一覧	0.62	0.16	0.39

表 4　伏見稲荷大社のページ集合における，各コサイン類似度と Score

ページタイトル	属性「歴史」	八坂神社の歴史	Score
全国稲荷会	0.51	0.16	0.34
ご祭神	0.48	0.18	0.33
講務本庁	0.46	0.18	0.32
伏見稲荷大社とは	0.47	0.09	0.28
よくあるご質問	0.45	0.08	0.26

ページタイトル	属性「祭神」	御祭神	Score
ご祭神	0.50	0.18	0.34
ご祈祷	0.30	0.11	0.20
よくあるご質問	0.26	0.06	0.16
伏見稲荷大社とは	0.20	0.06	0.13
祭礼と行事	0.15	0.05	0.10

ページタイトル	属性「現地情報」	境内マップ・アクセス	Score
交通アクセス	0.28	0.08	0.18
トピックス	0.28	0.07	0.17
伏見稲荷大社とは	0.31	0.03	0.17
おいなりさんへ	0.26	0.08	0.17
周辺マップ	0.23	0.08	0.15

ページは，それぞれ，表 3，表 4，表 5 における下線部のページタイトルである．3 つの適応オブジェクトにおける 3 つの内容の計 9 ページのうち，6 ページに関しては正解を上位 5 位以内に判定できているため，高い精度を得たと考えられる．よって，

表 5　東福寺のページ集合における，各コサイン類似度と Score

ページタイトル	属性「歴史」	八坂神社の歴史	Score
東福寺の歴史について	0.75	0.25	0.50
東福寺紅葉状況	0.47	0.19	0.33
東福寺駐車場閉鎖のご案内	0.46	0.20	0.33
記念講演と特別拝観	0.46	0.19	0.33
観光バスと一般車両の駐車	0.39	0.18	0.28

ページタイトル	属性「複合」	御祭神	Score
東福寺の歴史について	0.62	0.01	0.32
国指定名勝 東福寺本坊庭園	0.42	0.01	0.21
東福寺紅葉状況	0.42	0.00	0.21
東福寺駐車場閉鎖のご案内	0.40	0.00	0.20
記念講演と特別拝観	0.39	0.00	0.19

ページタイトル	属性「複合」	境内マップ・アクセス	Score
東福寺の歴史について	0.62	0.03	0.33
東福寺紅葉状況	0.42	0.01	0.21
国指定名勝 東福寺本坊庭園	0.42	0.01	0.21
東福寺駐車場閉鎖のご案内	0.40	0.01	0.20
記念講演と特別拝観	0.39	0.01	0.20

対応 Web ページ特定手法が有用であることを確認した．以下，9 ページの各順位に関して考察する．

京都タワーについて以下に述べる．八坂神社の属性名と一致するものは存在しなかったため，全て属性「複合」であり，スコアは全て同一のものである.「八坂神社の歴史」と「境内マップ・アクセス」に対応するページは，それぞれ，1 位の「タワーについて」と，2 位の「交通アクセス」となった．これらの内容は属性の一致やコンテンツの類似という指標が成功したと考えられる．八坂神社の「御祭神」は神社固有の属性であるため，存在しないという結果は正しいと考えられる．そのため，固有の属性に関する，対応するページの特定内容は今後検討する必要がある．

伏見稲荷大社について以下に述べる．まず,「八坂神社の歴史」に対応するページは，3 位の「講務本庁」となった．1 位の「全国稲荷会」は，その総会が開催された日時の記録表があり　，歴史で頻出する年表の名詞と大きく類似し，さらに，神社名と会名に共通する「稲荷」が多く含まれるために，Score が高くなったと考えられる．2 位の「ご祭神」は，歴史的背景と考えられる記述があるために，Score が高くなったと考えられる．次に，八坂神社の「御祭神」に対応するページは，1 位の「ご祭神」となった．八坂神社と同種のオブジェクトであるため，固有の属性であっても祭神を高い精度で抽出したと考えられる．最後に，八坂神社の「境内マップ・アクセス」に対応するページは，5 位の「周辺マップ」となった．システムにお

いて最もスコアが高くなった「交通アクセス」のページは，時期ごとの案内経路へのリンクが貼られているのみという構造であった．それに対して,「周辺マップ」は地図画像が表示されているため，被験者はより地理的な情報に近いものであると判断したと考えられる．本手法はテキスト処理がベースであるため，地図や画像は考慮してない．しかし，一般に，アクセス情報においてテキストだけでなく，地図や画像は視覚的に理解を補助すると考えられる．よって，アクセス情報の対応付けの際，地図が検出された場合は重みを付けることが必要であると考えられる．

東福寺について以下に述べる．まず,「八坂神社の歴史」に対応するページは，1 位の「東福寺の歴史について」となった．歴史は，社寺などの多くの種類のオブジェクトで使われる属性であるため，Score が高くなったと考えられる．次に，八坂神社の「御祭神」に対応するページは，京都タワー同様，存在しないという結果は正しいと考えられる．最後に，八坂神社の「境内マップ・アクセス」に対応するページは，存在しないと順位外の 2 つの結果が同列となった．本来のアクセスのページはトップからの 1 リンク以内に存在しなかったため，存在しないという結果は正しいと考えられる．順位外のページ「観光バスと一般車両の駐車について」が選択されたことに関しては，バスやアクセスといった単語が含まれるためであると考えられる．

全体を通して，属性「複合」を利用する内容が複数存在する場合，上位に現れる内容がほぼ同一となっている．また,「複合」は「歴史」の代替としては高い精度を得られたが，それ以外の属性に関しては低い精度となった．この原因として，属性「複合」の単語ベクトルは，様々な内容を含むことで「歴史」や「概要」と近いものとなっていることが考えられる．よって，属性「複合」を利用する場合においては重みを減らし，コンテンツの類似度をより強く反映させるべきであると考えられる．

6. まとめと今後の課題

ユーザが，ある Web サイトにおける概要やアクセス情報などの Web ページの一部を保存した場合，他オブジェクトの Web サイトにおいて，保存した Web ページの一部に対応する Web ページを特定する手法を提案した．実験において，Wikipedia を用いた属性抽出と対応 Web ページ特定手法の有用性を確認した．今後の課題として，ユーザの保存した内容とシステムの保存する内容の類似度と，システムによって判定された項目と保存する内容の類似度のバランスについて検討すべきである．具体的には，保存したテキストの分量，Wikipedia のテキストの分量，対応する Web ページのテキストの分量によって，コンテンツの類似度と属性の一致度の信頼性は変化すると考えられる．これらの性質や関係性を分析し，自動制御する方法について検討する必要がある．また，八坂神社と京都タワーのような種類が異なるオブジェクトに関して，祭神のような種類に依存するような情報の対応について検討が必要である．また，システムによって適応するサイト内のページ集合に関しても妥当な範囲を予備実験によって定める必要があると考えられる．また，プロトタイプシステムを用いて，被験者によって手動で必要なデータを保存した場合とシステムを利用した場合の所要時間を比較することで，手軽な情報収集の実現に関して評価を行う必要がある．

謝　　辞

本研究の一部は，平成 26 年度科研費基盤研究 (B)(課題番号 : 26280042) ならびに平成 26 年度科研費若手研究 (B)(課題番号 : 24700098) によるものです．ここに記して謝意を表すものとします．

文　　献

[1] 峯祥平，北山大輔： スマートスクラップブックのための観光情報収集時における意図抽出に関する一考察，電子情報通信学会技術研究報告．DE, データ工学，Vol.114, No.101, pp.83-86, 2014.

[2] 小谷彬，大島裕明，小山聡，田中克己: 複数 Web サイトからの共通属性抽出による共通サイトマップの生成，電子情報通信学会技術研究報告. DE, データ工学，Vol.106, No.149, pp.35-40,2006

[3] 佃洸摂，大島裕明，加藤誠，田中克己: オブジェクト間の意外な共通点の発見，DEIM Forum 2014，A8-1，2014

[4] 田中祥太郎，ヤトフトアダム，田中克己: ニュース記事の理解支援のための背景知識抽出と補完，情報処理学会研究報告. 情報学基礎研究会報告，Vol.2014-DBS-159 No.17

[5] 加藤 誠, 大島 裕明, 田中 克己: 例示検索のための集約点に基づくドメイン適応，日本データベース学会論文誌 DBSJ Journal 11(1), pp.49-54,2012

[6] 三笠弘貴，奥野拓: 観光サイトにおける閲覧目的に基づいた旅行記概要の動的生成，情報処理学会研究報告. DD，Vol.2014, No.4, pp.1-8，2014

[7] 石野亜耶，小林大祐，難波英嗣，竹澤寿幸: ブログを利用した観光情報リンク集の自動構築， 言語処理学会 第 16 回年次大会, PP246-249, 2010

[8] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237, 2004

[9] 北研二，津田和彦，獅々堀正幹: 情報検索アルゴリズム，共立出版，2002 年