

Wikipedia のリンク構造に基づく関連度を利用したコンテンツ検索手法と地球科学データへの応用

立床 雅司[†] 高橋 慧[†] 齊藤 昭則^{††} 吉川 正俊[†]

[†] 京都大学大学院 情報学研究科 〒 606-8501 京都市左京区吉田本町

^{††} 京都大学大学院 理学研究科 〒 606-8502 京都市左京区北白川追分町

E-mail: [†]{m.tatedoko, atakahashi}@db.soc.i.kyoto-u.ac.jp, ^{††}saitoua@kugi.kyoto-u.ac.jp,
^{†††}yoshikawa@i.kyoto-u.ac.jp

あらまし 本研究では、写真や動画、地球科学データなどといった時間情報、空間情報を持つコンテンツに対するキーワード検索を行うために、Wikipedia のリンク構造に基づく関連度を利用したコンテンツ検索手法を提案する。Wikipedia のリンク構造を用いることにより、記事と記事のつながりの強さ、記事と時間情報のつながりの強さ、記事と空間情報のつながりの強さを抽出する。コンテンツは最も類似している記事と結びつける。クエリと最も類似している記事を基点記事とし、基点記事からたどることのできる記事に結び付けられたコンテンツに対して、基点記事から抽出できる時間情報、空間情報を用いてランキングを行う。また基点記事と他の記事のつながりの強さを用いて関連トピックのランキングを行う。提案手法の応用例として、地球科学データ検索への適用を行う。

キーワード e-Science, KML, Wikipedia, リンク構造

A Method of Data Retrieval Based on the Link Structure of Wikipedia and its application to Earth Science data

Masashi TATEDOKO[†], Akira TAKAHASHI[†], Akinori SAITO^{††}, and Masatoshi YOSHIKAWA[†]

[†] Graduate School of Informatics, Kyoto University

Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501 Japan

^{††} Graduate School of Science, Kyoto University

Kitashirakawa-Oiwakecho, Sakyo-ku, Kyoto, 606-8502 Japan

E-mail: [†]{m.tatedoko, atakahashi}@db.soc.i.kyoto-u.ac.jp, ^{††}saitoua@kugi.kyoto-u.ac.jp,
^{†††}yoshikawa@i.kyoto-u.ac.jp

Abstract In this paper, we propose keyword search for contents retrieval using relationship based on the link structure of wikipedia. We extract strength of the connection of article and article and strength of the connection of article and spatial information and strength of the connection of article and temporal information from wikipedia. Contents are associated with the most similar article. We apply proposal to earth science data retrieval.

Key words e-Science, KML, Wikipedia, link-structure

1. はじめに

近年 Web 上において写真や動画、地球科学データといった多種多様なコンテンツが公開されている。これらのコンテンツの共通点は、写真や動画であれば撮影日や GPS の位置情報、地球科学データであれば観測期間や観測場所といった時間と空間に結びついたデータであるといえる。このようなコンテンツを検索するために、我々は Wikipedia を知識ベースとして用いたキーワード検索手法を提案する。時間情報と空間情報を持った

コンテンツに対する検索として時間範囲と空間範囲を指定することが有用であると考えられる。この場合、検索の対象とするコンテンツをあらかじめ選ぶ必要があるが、膨大なコンテンツに対しては困難である。特に地球科学データのような専門性の高いコンテンツではどのデータセットにどのような情報が含まれているのかを判断することは地球科学の研究者であっても困難である。たとえ地球科学の研究者であっても、自分の専門分野以外については知識がないからである。地球温暖化の研究のような新しい分野の研究のためには、分野をまたいだ多様なデー

タの比較が必要であるため、専門分野以外の地球科学データを検索したいというニーズがある。それゆえ、データセットの選択のために専門語を用いないキーワード検索が必要であると考えられる。また、専門語を用いないキーワード検索ができれば一般の人も地球科学データに触れることができるようになる。地球科学データは一般の生活にも結びついているため、簡単な検索によって天気以外のデータも活用される可能性が出てくる。

本研究では、時間情報や空間情報と結び付けられたコンテンツを簡単なキーワードで検索できるようにするため、Wikipediaから抽出した知識ベースを用いた検索手法を提案する。Wikipediaから抽出する知識は、トピック同士のつながりの強さとトピックに結び付けられた時間情報、空間情報のつながりの強さである。トピック同士のつながりの強さを求めるために既存の手法を用いる。トピックと時間情報、空間情報との結び付きの強さを抽出するために提案手法を用いる。

本論文では、2章で提案手法の関連研究について述べる。3章では、Wikipediaのリンク構造に基づく関連度を利用したコンテンツ検索手法を提案し、4章で地球科学データ検索への応用について述べる。5章では評価に向けた議論を行い、6章でまとめと今後の課題を述べる。

2. 関連研究

Google EarthとKMLを用いた研究のアウトリーチ手法としてDagik, Google Earth Outreach [7], NSIDCによるMapping Glacier Data and Photographs via GeoServer and Virtual Globes [8]などが挙げられる。本研究では、検索結果の出力先としてDagikを利用する。Dagikはすべてのデータセットへのリンクを持つKMLファイルをダウンロードさせ、Google Earth上でデータセットや期間を選択させる手法をとっている。このため、図1に示すように、複数の分野のデータセットを同時にプロットし比較することができる。Google Earth Outreach

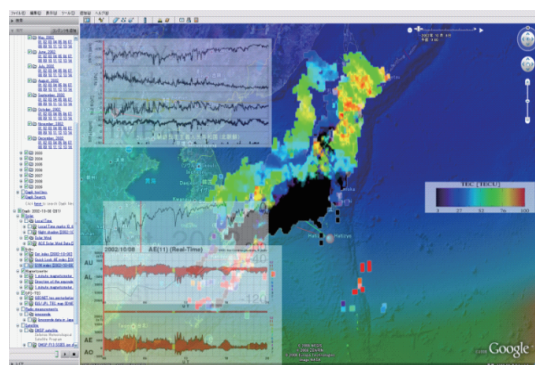


図1 Dagikによる複数データプロットの画面

やNSIDCによる手法では、1つのデータセットに対して1つあるいは複数のKMLファイルが対応している。これらの手法には検索が行えないという問題点がある。Dagikではタイトルだけでデータセットの種類を判断しなければならず、Google Earth OutreachやNSIDCによる手法でも同様にタイトルや説明文を1つ1つ見ていかなければならない。このため、専門

的な知識を持たない利用者にとってデータセットの発見は困難である。本研究では、Dagikに登録されているデータセットを一般的なキーワードで検索しランキングする手法を提案する。

地球科学データのポータルサイトとしては、NASAによるGlobal Change Master Directory (GCMD) [2]やLEAD (Linked Environments for ATMOSPHERIC Discovery) [4]やGEON [3]などが挙げられる。GCMDには、国内外の研究機関による多数のデータセットが登録されており、データセットの検索、概要閲覧、実データへのリンクを提供している。データセットの検索は、メタデータの全文検索、メタデータの項目に対する全文検索が可能である。検索の追加項目として、図2に示すように、地図を用いたユーザインタフェースによる緯度経度の指定や、開始年月日と終了年月日の指定を行うことができる。メタデータへの全文検索は単純な文字列の



図2 GCMDの検索画面

一致のみであり、オーロラの発光と地磁気の乱れに関連があるという知識を持ったユーザでなければ、地磁気に関するデータセットを検索することができない。地球科学データのクイック・ルックに適したシステムではなく、専門家が自身の専門分野のデータセットを検索する場合のみに使えるといえる。LEADやGEONといった実験的なポータルも全文検索やオントロジーを用いた入力支援、空間範囲指定、時間範囲指定といった検索が行えるが、専門知識がなければ検索できないという同様の問題がある。以上のように現在存在する地球科学データのポータルサイトでは、専門的な知識を持ったユーザのみを対象としており、専門分野以外の地球科学データを探したいユーザや専門知識を持たないユーザは対象外である。本研究では、Wikipediaを用いたオントロジーにより「オーロラ → 地磁気」といった知識を提供し、専門的な知識を持たないユーザにも地球科学データに触れる機会を与える手法を提案する。

地球科学の分野では、オントロジー作成の試みとしてNASAによるSWEET (Semantic Web for Earth and Environmental Terminology) が挙げられる。SWEETは地球科学データの概念を体系化するものであり、汎用的な語彙とのつながりをサポートしていない。本研究では、専門的な用語と汎用的な語彙とのつながりを表現するために、一般的な知識源であるWikipediaを利用する。

Wikipedia を利用したオントロジーの構築研究として、YAGO [9] が挙げられる。YAGO は WordNet と Wikipedia のページをマッチングさせた巨大な Web オントロジーである。本研究では、Wikipedia の記事間のつながりの強さを求めるため、リンク構造を用いた関連度の数値化を行う。

3. Wikipedia のリンク構造に基づく関連度を利用したコンテンツ検索手法

3.1 概要

本章では、Wikipedia のリンク構造に基づく関連度を利用したコンテンツの検索手法を提案する。対象となるコンテンツは時間情報、空間情報を持つものとする。本手法の概要を図 3 に示す。コンテンツの検索は、次の手順で行う。

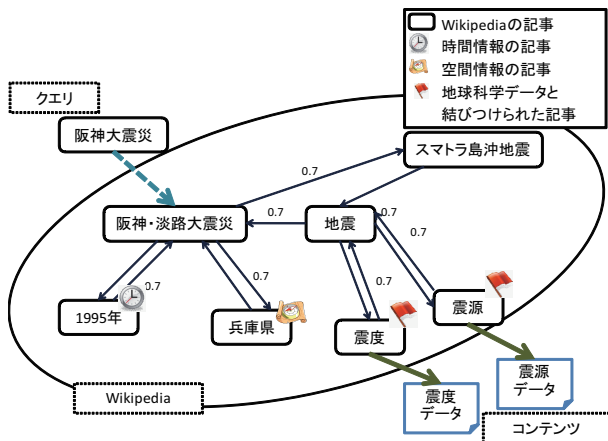


図 3 Wikipedia のリンク構造に基づく関連度を利用したコンテンツの検索手法の概要

- (1) クエリを Wikipedia の検索エンジンで評価し、最上位の結果を基点記事とする、
- (2) 基点記事のリンクをたどり、2 ホップ先までの記事を取得し、それぞれの記事間のリンクの重みを求める、
- (3) 基点記事にとって重要な時間情報、空間情報を抽出する、
- (4) (2) で求めたリンクを用いて、基点記事と結びつけられているコンテンツを取得する、
- (5) (4) で得られたコンテンツについて、(2) で得られた記事間のリンクの重みと (3) で得られた時間情報、空間情報を用いてランキングを行う、
- (6) 評価結果として基点記事に関連するコンテンツのランキングと関連記事のランキングを返す。

図 3 の例では、「阪神大震災」というクエリに対して「阪神・淡路大震災」が基点記事となる。基点記事に関連のあるコンテンツとして震度データ、震源データが得られる。基点記事から時間情報、空間情報の関連度を求め、これらと記事間関連度を用いてコンテンツをランキングする。関連記事は、スマトラ島沖地震などが記事間関連度によりランキングされる。検索結果は以下ようになる。

- クエリ: 阪神大震災

- 基点記事: 阪神・淡路大震災
- コンテンツ:
 - 震度 1995 年 1 月 17 日
 - 震源 1995 年 1 月 17 日
 - ...
- 関連記事
 - スマトラ島沖地震
 - ...

3.2 基点記事の取得

基点記事は、Wikipedia の検索エンジンに依存している。クエリと一致するタイトルを持つ記事がある場合はその記事が基点記事となる。そうでない場合は、Wikipedia の検索エンジンにおいて最上位の検索結果を基点記事とする。いずれの場合も、リダイレクトされる場合はリダイレクト先の記事を基点記事とする。

クエリ q から得られる基点記事 a は次のように定義する。

$$b(q) = a \quad (1)$$

3.3 Wikipedia の記事間の関連度

Wikipedia の記事間のリンクの重みを取得するために既存の手法を用いる。ここでは、中山らが提案した $pfibf$ [6] を用いる。 $pfibf$ (Path Frequency Inversed Backward link Frequency) は、Wikipedia のページを v_i, v_j , v_i から v_j への n ホップ先への全経路を $T = \{t_1, t_2, \dots, t_n\}$ とすると、以下の式で表される。

$$pfibf(v_i, v_j) = \sum_{k=1}^n \frac{1}{d(|t_k|)} \cdot \log \frac{N}{bf(v_j)} \quad (2)$$

ここで、 d は経路長に応じて単調増加する関数であり、 N は全記事数、 $bf(v_j)$ は他のページからのリンク数である。

本手法では、基点記事から 2 ホップ先までの記事の関連度を求める。記事 a_i, a_j 間の関連度は次のように定義する。

$$Ra(a_i, a_j) = pfibf(a_i, a_j) \quad (3)$$

3.4 記事に関連する時間情報、空間情報の抽出

本手法では、次のような仮定を基に記事からの時間情報、空間情報を抽出を行う。

- (1) 記事に関連する時間情報、空間情報とは、記事からリンクされている 1 ホップ先の記事の属性とする。記事からリンクのある記事を有用なものと思なすためである、
- (2) リンク先記事からのバックリンクがある場合、記事との結び付きの高い時間、空間情報であると定義する。そのため、バックリンクの重みを行きのリンクよりも大きくする、
- (3) 記事の属性は記事のカテゴリで判別する。時間情報の場合は、記事のカテゴリからカテゴリ「暦法」へのパスがあるとき、空間情報の場合は、記事のカテゴリからカテゴリ「場所」へのパスがあるときにそれぞれの属性が付与される。

記事を a , 属性を持つ記事を a_1 , 記事からのリンクの重みを定数 u ($0 < u < 1$) とし、 a, a_1 間の関連度を以下のように定義する。

$$r(a, a_1) = \begin{cases} 1 & \text{相互リンクを持つ} \\ u & \text{記事からのリンクのみ} \end{cases} \quad (4)$$

以下では、時間情報、空間情報それぞれに特化した情報抽出手法を述べる。

3.4.1 記事に関連する時間情報の抽出

本手法では、Wikipedia の時間情報に関する記事のうち、年の記事と月日の記事とを対象とする。年と月日の記事の結びつきの強さを表現するために、記事に出現するリンクの順序の近さを用いる。基点記事を a 、基点記事からリンクしている年の記事を t_{y_i} 、基点記事からリンクしている月日の記事を t_{d_j} とすると、記事に関連する時間情報の関連度 tr は次のようになる。

$$tr(a, t_{y_i}, t_{d_j}) = t_{y_i} \cdot t_{d_j} \cdot d(t_{y_i}, t_{d_j}) \cdot m(t_{y_i}, t_{d_j}) \quad (5)$$

ここで、 $order(a, t)$ は記事 a 内の t へのリンクの順序、 $order_count(a)$ は記事 a 内のリンクの総数とすると、

$$d(t_{y_i}, t_{d_j}) = \begin{cases} 1 - \frac{order(a, t_{d_j}) - order(a, t_{y_i}) - 1}{order_count(a)} & \text{if } order(a, t_{d_j}) - order(a, t_{y_i}) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

であり、

$$m(t_{y_i}, t_{d_j}) = \begin{cases} 1 & \text{if } r(t_{y_i}, t_{d_j}) = 1 \wedge r(t_{y_i}, t_{d_j}) = 1 \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

である。

3.4.2 記事に関連する空間情報の抽出

空間情報も時間情報と同様の手法で抽出する。基点記事を a 、基点記事からリンクしている場所の記事を s_i とすると、記事に関連する空間情報の関連度 sr は次のようになる。

$$sr(a, s_i) = r(a, s_i) \quad (8)$$

3.5 記事とコンテンツとの結びつけ

各コンテンツは 1 つの記事に結びついている。コンテンツと記事を結びつける手法は、コンテンツの内容に依存する。Wikipedia の記事はテキストデータを持っているため、tfidf によるコンテンツとの類似度計算などにより算出することが考えられる。

3.6 記事とコンテンツの関連度

コンテンツを c_i とすると、記事とコンテンツの関連度は次のように定義できる。

$$Rc(a, c_i) = Rc_t(a, c_i) \cdot Rc_s(a, c_i) \quad (9)$$

$$Rc_t(a, c_i) = \begin{cases} 1 & \text{if } c \text{ の期間が } a \text{ の持つ時間情報を含む} \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

$$Rc_s(a, c_i) = \begin{cases} 1 & \text{if } c \text{ の空間が } a \text{ の持つ空間情報を内包} \\ 0.3 & \text{if } c \text{ の空間が } a \text{ の持つ空間情報と交差} \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

3.7 クエリとコンテンツの関連度

以上から、クエリとコンテンツの関連度は次のように定義できる。

$$R(q, c_i) = Rc(b(q), c_i) \quad (12)$$

4. 地球科学データへの応用

4.1 概要

本章では、Wikipedia のリンク構造に基づく関連度を利用したコンテンツ検索手法の応用例として地球科学データの検索への適用を行う。検索対象となるコンテンツは、Dagik がリストを提供している地球科学データである。Dagik には現在衛星太陽風データや地磁気 SYM 指数など 8 分野 29 種類のデータセットが登録されている。地球科学データは KML 形式で表現されている。地球科学データ検索システムの概要を図 4 に示す。

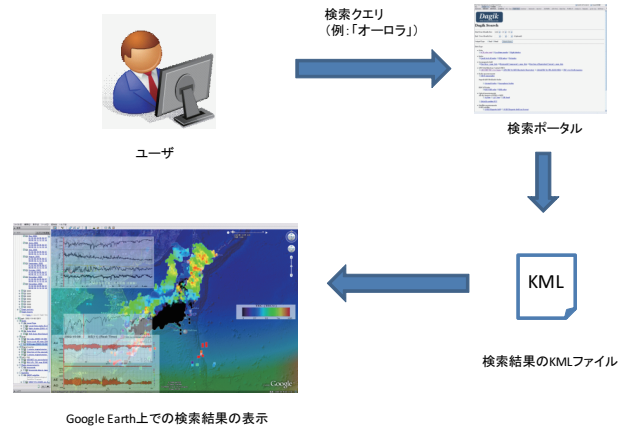


図 4 地球科学データ検索システムの概要

地球科学データ検索システムの入力、ユーザーから入力されたキーワードである。キーワードをクエリとしてコンテンツ検索手法を適用し、ランキングされた KML 文書のリストを取得する。システムは出力として KML 文書のリストをマージした KML 文書をユーザーに返す。ユーザーは KML 文書をダウンロードし、Google Earth に読み込むことにより、プロットされたデータを閲覧できる。

4.1.1 想定するユーザ

提案システムでは、専門分野以外の地球科学データをクイック・ルックしたいユーザを対象とする。地球科学データの研究者でも、自身の専門分野から少し離れるだけでどのようなデータセットが存在するのか、どこにデータがあるのかを容易に把握することができなくなる。また、地球科学データを専門としない利用者にとっては、地球科学にどのような分野があるのかを把握することも困難である。

このようなユーザの地球科学データに対するクイック・ルッ

クというニーズに対する検索支援として、自身の専門分野に関する他分野の地球科学データや身近なキーワードに対する地球科学データの検索機能を提供する。キーワードの入力は日本語と英語を受け付ける。

4.1.2 利用環境

提案システムは、Web 上のポータルサイトとして提供される。検索結果は、Google Earth に表示される。ユーザは、Web ブラウザと Google Earth が起動でき、インターネットに接続できる環境があればよい。Web を利用する利点は、Web ブラウザさえあれば追加のインストールなしで新しいアプリケーションが利用できることである。現在の手法では、Google Earth という追加のアプリケーションが必要であるが、Google Earth をブラウザ上で実行するプラグインの開発が進められており、将来的にはブラウザの中で閉じた環境を構築することも視野に入れている。Google Earth のブラウザプラグインの現在のバージョンでは時間をコントロールする機能が実装されていないため、本提案では利用していない。

4.2 KML ファイル

KML(Keyhole Markup Language) は、地理空間情報を記述するための XML ベースのマークアップ言語である。KML は Google Earth の前身を開発した Keyhole 社独自の言語であったが、2008 年に地理情報システム関連の国際標準化団体 Open Geospatial Consortium(OGC) から標準化の承認を受けた。KML は Google Earth の他、NASA の World Wind、Microsoft の Virtual Earth などのデジタル地球儀でも利用できる。KML では、目印やパス、オーバーレイなどの位置指定を緯度経度の指定により行う。

Dagik で扱われている KML ファイルは、1 ファイルに 1 つのデータセット 1 日分が記述され、データセットの配布元へのリンクが関連付けられている。提案手法では、KML ファイルに記述された情報のうち、データセットの配布元の Web ページ、時間情報、空間情報を利用する。Dagik の KML ファイルの例を図 5、図 6 に示す。図 5 はデータセットの配布元へのリンクの例である。データセットの説明とリンクが **description** 要素内に記述されている。

```
1 <NetworkLink>
2   <visibility>0</visibility>
3   <open>0</open>
4   <name>Infrared window cloud image (ISCCP B1 data)
5     [2007-01-01]</name>
6   <description>
7     <![CDATA[ <P> International Satellite Cloud
8       Climatology Project (ISCCP) B1 data for cloud
9       image is provided by <font color="#50c0ff"> A
10      HREF="http://www.ncdc.noaa.gov/oa/rsad/
11      isccpb1/index.php">NOAA/NCDC</font>. <P> <
12      HR> Plotted by Dagik team of Kyoto University/
13      SPEL. <P> ]]>
14   </description>
15   <styleUrl>#GEONET_map_BalloonStyle</styleUrl>
16   <Snippet maxLines="1"></Snippet>
17   <Url>
18     <href>http://www-step.kugi.kyoto-u.ac.jp/dagik/data/
19       clouds/2007/clouds_IR_2007_01_01.kmz</href>
20   </Url>
21 </NetworkLink>
```

図 5 データセットの配布元へのリンクの例

図 6 は時間情報、空間情報の例である。GroundOverlay 要素の子の TimeSpan 要素に時間情報が記述され、LatLonBox 要素に空間情報が記述されている。

```
1 <GroundOverlay>
2   <name>03_00_00</name>
3   <TimeSpan>
4     <begin>2007-01-01T03:00:00Z</begin>
5     <end>2007-01-01T05:59:59Z</end>
6   </TimeSpan>
7   <color>CCFFFFFF</color>
8   <Icon>
9     <href>./images/2007-01-01_03_00_00.jpg</href>
10  </Icon>
11  <altitude>10000.0</altitude>
12  <altitudeMode>absolute</altitudeMode>
13  <LatLonBox>
14    <south>-60.0000</south>
15    <north>60.0000</north>
16    <west>-180.000</west>
17    <east>180.000</east>
18    <rotation>0.</rotation>
19  </LatLonBox>
20 </GroundOverlay>
```

図 6 時空間情報の記述例

5. まとめと今後の課題

本研究では、Wikipedia のリンク構造に基づく関連度を利用したコンテンツ検索手法を提案し、地球科学データへ応用したシステム試作のための考察を行った。Wikipedia の記事から時間情報、空間情報を抽出するための手法を提案し、コンテンツをランキングするための関連度の定義を行った。

今後、時間情報、空間情報を抽出する提案手法について評価を行う必要がある。その評価を基に、コンテンツ検索手法をより洗練させていきたい。地球科学データへの応用例に対しては、精度、再現率による定量評価を検討している。

文 献

- [1] Dagik (Daily Geospace data in kml), <http://www-step.kugi.kyoto-u.ac.jp/dagik/>
- [2] Global Change Mastaer Directory Discover Earth science data and services, <http://gcmd.nasa.gov/KeywordSearch/Home.do>
- [3] GEON (Geosciences Network), <http://www.geongrid.org/>
- [4] LEAD (Linked Environments for Atmospheric Discovery), <http://lead.ou.edu/>
- [5] SWEET, <http://sweet.jpl.nasa.gov/>
- [6] K. Nakayama, T. Hara and S. Nishio: “Wikipedia mining for an association web thesaurus construction.”, Proc. of IEEE International Conference on Web Information Systems Engineering (WISE 2007) (2007).
- [7] Google Earth Outreach, <http://earth.google.com/outreach/index.html>
- [8] National Snow and Ice Data Center, http://nsidc.org/data/virtual_globes/
- [9] Fabian M. Suchanek and Gjergji Kasneci and Gerhard Weikum: Yago: A Core of Semantic Knowledge, 16th international World Wide Web conference (WWW 2007).