リンク解析を用いた地球環境ポータルサイト構築の試み

豊田 正史
科学技術振興事業団
計算科学技術活用型特定研究開発推進事業
連絡先: 東京大学生産技術研究所
喜連川研究室
mtoyoda@acm.org

菊地 時夫
高知大学 理学部 数理情報科学科
tkikuchi@is.kochi-u.ac.jp

近年、地球環境問題の重要性が増すにしたがって、地球環境に関する web ページも増加を続けている。地球環境に関する話題は、気象、海洋、地質など多岐に渡っており、関連する組織も、政府機関から、大学、企業、民間公益団体 (NGO) など幅広い。このため、地球環境情報を提供する web ページへのリンクを集めたポータルサイトを構築することは、地球環境の研究者や専門家にとって有用である。我々は、リンク解析を用いた web ページの分類および関連付け手法を開発しており、本論文ではこの手法がポータル作成者にとって有効なツールとして働くかを確認する調査を行った。

1 はじめに

近年、地球環境問題の重要性が増すにしたがって、地球環境に関する web ページも増加を続けている。地球環境に関する話題は、気象、海洋、地質、農業、水環境など多岐に渡っており、関連する組織も、政府機関から、研究機関、企業、民間公益団体 (NGO) など幅広い。地球環境の研究者にとって、これらのページから得られる情報は重要であるが、どこにどのような情報があるのか把握するのは難しい。このため、地球環境情報を提供する web ページへのリンクを集めたポータルサイトを構築することは有用である。NASA では、Earth Science Portal[1] という地球科学の研究に関するポータル構築が行われているが、日本においてはまだこういった試みは行われていない。また、我々は研究に関するページのみでなく、より幅広い範囲のポータル構築を目指している。

ポータルの作成は、最終的には人手に頼らざるを得ない。しかし、HITS[2] を代表とするリンク解析手法は、ポータル作成の有効なツールとなる可能性がある。HITS は、web のグラフから同じトピックを扱う web ページの集合 (例えば、気象情報を提供するページの集合) を抽出する手法である。こうしたページの集合を、コミュニティと呼ぶ。抽出されたコミュニティは、ポータルサイトにおけるカテゴリとして利用できる可能性がある。

我々は、[3] において、与えられた web ページから関連するコミュニティ群を発見する手法を提案している。本論文では、この手法 [3] を地球環境関係のページに適用し、これがポータル作成にとって有効なツールとして働くかどうかを確認する調査を行った。

2 関連コミュニティ群発見手法の適用

我々が [3] で提案した手法は、ある web ページ s をシードとして与えると、s を含むコミュニティに加えて、s に関連するコミュニティを幾つか出力するものである。したがって、この手法に様々な地球環境に関するシードを与えることで、関連する多くのコミュニティを得ることができる。また、シード s を含むコミュニティと、s から導出されたコミュニティが関連していることを利用して、コミュニティ間の関連を調べることもできる。以下に、この利用方法の概要を述べる。

1. 地球環境に関係するページを集め、シードセットを用意する。
2. シードセット中の各シードに付いて別々に、[3]

の手法を適用する。結果、各シードについて幾つかのコミュニティが得られる。

3. ステップ 2 で得られたコミュニティのうち類似するコミュニティを 1 つにまとめる。類似度は、コミュニティ間で共有するページの数などを基に決定する。
4. シードを含むコミュニティと、シードから導出されたコミュニティを関連付ける。

本論文では得られた結果の調査に重点をおくため、ステップ 2,3,4 の詳細に付いては別途論文として発表する予定である。ステップ 2 で利用している関連コミュニティ群発見手法の詳細に付いては、[3] を参照されたい。

この手順により、シードセットに関連したコミュニティの集合が得られる。以下では、(1) これらのコミュニティを、ポータルサイトのカテゴリとして利用できるかどうか、(2) コミュニティ間の関連をカテゴリ間のリンクとして利用できるかどうかを、実験により確認する。

3 地球環境シードセットを用いた実験と結果

この実験では、Web ロボットを使って収集した約 1400 万の web ページを、データセットとして用いた。ロボットは、jp ドメインにあるページ、または日本語で書かれたページを収集するようになっている。収集は、2000 年の 7 月から 8 月にかけて行った。

シードセットとして、気象および水資源に関係して、研究および情報提供などを行っているページの URL、約 80 個を用いた。シードの大部分は高知大学の気象情報リンク (http://weather.is.kochi-u.ac.jp/links.html) にリストアップされている。

このシードセットに対して、第 2 節の方法を適用した結果、65 のコミュニティが得られた。コミュニティの完全なリストは、http://www.tkl.iis.u-tokyo.ac.jp/~toyoda/weather/communities.html で見ることができる。以下にコミュニティの具体例を 2 つ示す。

日本天気電気学会
日本海洋学会
日本地震学会
日本気象学会
日本雪氷学会
日本水文科学学会
日本地質学会
日本地理学会
日本火山学会

IMOC Weather Page
Kochi University Weather Home
新日本気象海洋株式会社
WeatherEye
防災気象情報サービス
ヤン坊マー坊天気予報
WNI Cyber Weather World
TBS WEATHER GUIDE

左は地球環境に関係する日本の学会が集まったコミュニティであり、右は日本の天気予報サービスのコミュニティである。

また、図 1 には、コミュニティの関連図の一部を示した。数字の入った矩形が、各コミュニティとその識別子を表しており、その間に引かれた線は、関連するコミュニティ同士を結んでいる。各コミュニティには、内容を表す説明が付加されているが、これに関しては、第 4 節で説明する。

4 評価

本節では、実験結果の定性的評価を行う。まず、コミュニティを説明するキーワードを決定し、その上で各コミュニティの分類、および関連付けに関する調査を行った。

4.1 キーワードによるコミュニティの説明

まず、コミュニティに含まれる全ページに、そのページを説明するキーワードを付け、その上で、各コミュニティを説明するキーワードの選定を行った。この結果、多くのコミュニティは表 1 に挙げたキーワードで説明できた。キーワードは 4 種類あり、それぞれ活動分野、活動形態、ページ開設の主体、国を表す。これは様々なキーワードの分類を試みた中で、最も良くコミュニティの区別を説明できる組合せとなっている。

分野では、**地球環境**全般に関わるもの、**宇宙・地理**といった周辺科学を第 1 レベルとし、それらの中で特定の領域に特化している場合には第 2 レベルのキーワードを付けた。コミュニティに第 2 レベルのキーワードが多数含まれる場合には、地球環境と一般化して説明する。地球環境に関係のないページが多数含まれる場合にはこのキーワードは空欄とする。この中で主に天気予報・天気図・気象衛星画像といった**一般気象**に関するものを特に分けて、研究中心の**大気**と区別した。2 番目のキーワード群は web ページを通じて行われている活動の方向性である。

3 番目にはページを開設している組織を表すキーワードを付けた。これも第 1 レベルとして大きく括った**国際機関・政府機関**などを挙げ、そのコミュニティを構成するページが特定の組織だけからなる場合には，その組織名を付けた[1]。コミュニティに、何種類

[1] **NOAA** は National Oceanic and Atmospheric Administration (米国海洋大気庁)、**NASA** は National Aeronautics and Space Administration (米国航空宇宙局)、**USGS** は

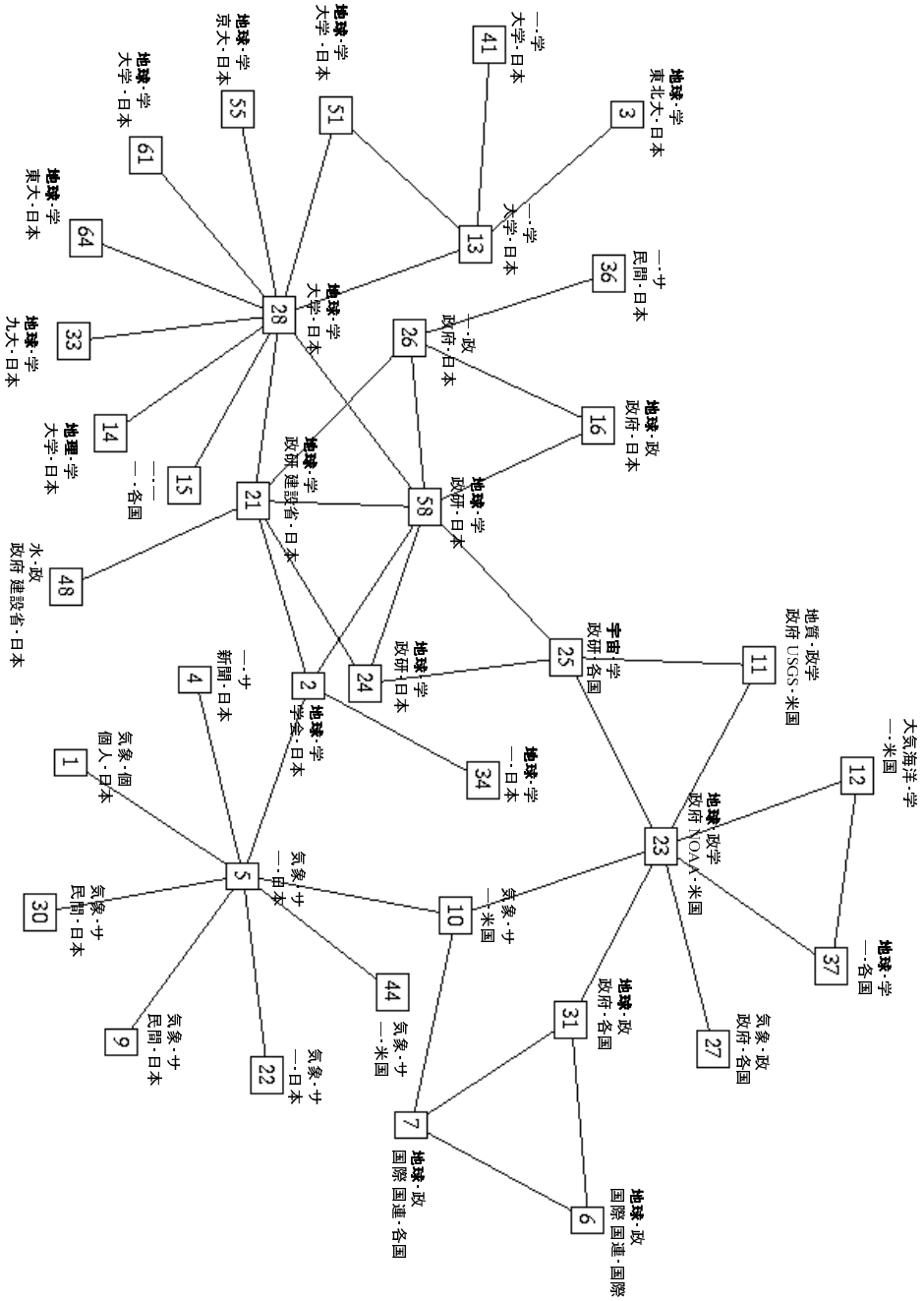

図 1: コミュニティ間の関連図

表 1: コミュニティを説明するキーワード

分野		活動形態	組織		国
第 1 レベル	第 2 レベル		第 1 レベル	第 2 レベル	
地球環境	大気	研究	国際機関	国連	国際
地理	気象	サービス	政府機関	NASA	日本
宇宙	海洋	政治	大学	NOAA	アメリカ
	水	私的活動	学会	USGS	各国
	雪氷		民間	科技庁	
	地質			建設省	
	植生			文部省	
	生物			ワシントン大学	
	防災			京都大学	
	衛星			九州大学	
				東大生産研	
				東北大学	
				北海道大学	
				新聞社	
				個人	

もの組織が含まれる場合、このキーワードは空欄とした。最後にそれぞれの組織の国籍を付加した。

表 2 には、65 のコミュニティにどのようなキーワードが付けられたかを示した。65 のコミュニティのうち、分野に関して地球環境に関連すると判断されたもの数は 47 個であった。残りの 18 のコミュニティは、トピックがより一般的になってしまったもの (例えば 13 番は大学のトップページの集まりである) および、まったく関係のないものなどである。

関連ないと判断されたコミュニティにも重要なものが存在する。例えば、13 番のような一般的なトピックのコミュニティは、多数のコミュニティをつなぐ糊の役割をするため (図 1 を参照)、コミュニティ間の関連を説明するのに必要な場合が多い。

4.2 各コミュニティの精度

次に、各コミュニティの中に適切なページがどの程度含まれているかを調べた。表 2 における精度の項に、コミュニティに含まれるページ数に対して、我々が主観でキーワードに合致すると判断したページの数を示している。地球環境とは関係のないコミュニティに関しては、この評価は行わなかった。表 2 に示されているように、各コミュニティの精度は非常に高い。地球環境に関連する 47 のコミュニティの内、キーワードに合致しないページを含んでいるものはわずか 6 個である。

United States Geological Survey (米国地質調査所) のそれぞれ略である。

4.3 分類の傾向

上述のキーワードによるコミュニティの説明が図 1 に示されている。右下に天気予報・気象情報サービス (5 番周辺) が分野によって集まっている他は、組織による分類が多く行われているのが分かる。中央には日本の政府機関および研究所が集まり (58 番周辺)、左側には大学が集まっている (28 番周辺)、また右上には国外のコミュニティが集まっている (23 番周辺)。

これらのコミュニティは、分野においてはほとんど第 1 レベルにおいて説明されており、分野第 2 レベルのキーワードのみで説明できるコミュニティは非常に少ない。これは、今回用いた地球環境のページに関しては、分野を細かく分類したリンク集より、組織による分類 (たとえば大学の学科単位) をしたリンク集の方が、数多く web 上に存在していることを示している。したがって、この例に関しては、天気予報のように一般の人々からよくリンクされているページ以外は、組織による分類が起こりやすい傾向

表 2: 各コミュニティのキーワードと精度

No.	精度 合致/個数	分野	活動形態	開設主体	国
1	7/7	気象	私的活動	個人	日本
2	9/9	地球環境	研究	学会	日本
3	5/5	地球環境	研究	大学:東北大学	日本
4	-/5	-	-	民間:新聞社	日本
5	8/8	気象	サービス	-	日本
6	4/4	地球環境	政治	国際機関:国連	国際
7	3/3	地球環境	政治	国際機関:国連	国際
8	3/3	水	研究	学会	各国
9	2/2	気象	サービス	民間	日本
10	3/4	気象	サービス	-	アメリカ
11	3/3	地質	研究	政府機関:USGS	アメリカ
12	4/4	大気:海洋	研究	-	アメリカ
13	-/6	-	研究	大学	日本
14	6/6	地理	研究	大学	日本
15	-/4	-	-	-	-
16	3/3	地球環境	政治	政府機関	日本
17	4/4	衛星	研究	大学	各国
18	2/4	地球環境	研究	民間	日本
19	3/3	海洋	研究	政府機関	各国
20	2/3	大気:海洋	研究	政府機関:大学	アメリカ
21	4/4	地球環境	研究	政府機関:建設省	日本
22	2/2	気象	サービス	-	日本
23	6/7	地球環境	研究:政治	政府機関:NOAA	アメリカ
24	3/3	地球環境	研究	政府機関	日本
25	7/9	宇宙	研究	政府機関	各国
26	-/6	-	政治	政府機関	日本
27	4/4	気象	政治:サービス	政府機関	各国
28	19/19	地球環境	研究	大学	日本
29	-/4	-	-	政府機関:NASA	アメリカ
30	3/3	気象	サービス	民間	日本
31	4/4	地球環境	政治	政府機関	各国
32	-/4	-	-	民間	日本
33	8/8	地球環境	研究	大学:九州大学	日本
34	3/3	地球環境	研究	-	日本
35	-/1	-	-	-	-
36	-/2	-	-	-	-
37	12/12	地球環境	研究	-	各国
38	-/3	-	-	-	-
39	2/2	雪氷	研究	大学	各国
40	2/2	植生	研究	-	各国
41	-/2	-	研究	大学	日本
42	-/3	-	-	-	-
43	2/2	地球環境	研究	政府機関:科技庁	日本
44	2/2	気象	サービス	民間	アメリカ
45	-/3	-	-	-	-
46	2/2	大気	研究	-	日本
47	-/2	-	研究	政府機関:文部省	日本
48	2/2	水	政治:サービス	政府機関:建設省	日本
49	1/1	気象	研究	大学	各国
50	-/2	-	-	-	-
51	7/7	地球環境	研究	大学	日本
52	-/2	-	-	大学:東北大学	日本
53	-/2	-	-	大学:北海道大学	日本
54	2/2	生物	サービス	民間	日本
55	2/2	地球環境	研究	大学:京都大学	日本
56	1/1	気象	サービス	政府機関	各国
57	-/1	-	-	-	-
58	4/4	地球環境	研究	政府機関	日本
59	-/2	-	-	-	-
60	2/2	水:雪氷	研究	大学:ワシントン大学	アメリカ
61	3/3	地球環境	研究	大学	日本
62	1/2	気象	研究	-	日本
63	2/2	大気:海洋	研究	大学:ワシントン大学	アメリカ
64	2/2	地球環境	研究	大学:東大生産研	日本
65	2/2	宇宙	研究	政府機関:NASA	アメリカ

表 3: 関連付けの種類と出現数

種類	出現数
分野の違い	6
分野の特定	2
組織の違い	14
組織の特化	5
分野も組織も同じ	6
分野も組織も違うが関連はある	13
説明不可能	2
合計	48

があると言える。この現象は、今回のシードセットが気象関係に偏っているために起きた可能性があり、様々な分野のシードを加えることで分類に変化が起きることも考えられる。

4.4 関連付けの傾向

組織による分類が多いため、コミュニティ間の関連の半分弱は、組織の違い、および特化で説明できる。例えば、以下のような例が挙げられる。

- 地球環境の研究をしているのは同じだが、大学か、政府研究所か、が異なる (28 番と 58 番)。
- 地球環境の研究をしているのは同じだが、大学が特定されていないか、京都大学の学科に特化されているか、が異なる (28 番と 55 番)。

また、分野の違いや特化で説明できる関連も存在する。例えば、28 番と 14 番は、地球環境一般の研究か、地理に関する研究かで区別が可能である。

表 3 には、図 1 に現れる関連付けの種類と出現数を示した。ほとんどすべての関連はなんらかの意味で説明可能であり、説明不能な関連は 2 つであった。組織の違いまたは特化で説明できるものが一番多く 19 あり、分野の違いまたは特化で説明できるものは 8 と比較的少なかった。また一方で、本来は同じコミュニティであるべきものを結ぶ関連も 6 つあり、これに関してはコミュニティ識別を改善することで対応する必要がある。

5 ポータルの構築

我々は、この結果を用いて実際にポータル構築を行った。このポータルは既に公開されており、現在 http://weather.is.kochi-u.ac.jp/ の「ポータル構築実験」というリンクから見ることができる。

我々が構築したポータルの特徴は、図 1 に示したマップをユーザのナビゲート用に利用している点にある。図 2 に、このポータルの画面スナップショットを示す。左側には、図 1 のマップがクリッカブルマップとして表示されている。分類が同じコミュニティが線で囲まれており、ユーザが分類を認識しやすくなっている。マップ内のコミュニティをクリックすることで、右側にページのリストが表示される。ページのリストの下には、関連するコミュニティのリストが関連度などと共に示されており、これはユーザがナビゲートする際の手がかりとなる。

6 まとめと今後の課題

本論文では、リンク解析手法を用いて地球環境ポータルサイトの構築を試みた。この結果、ある程度適切なページの分類が得られ、人手で最初からポータルを構築するよりも少ない手間でポータルを構築することができた。今回のシードセットは評価可能な範囲に絞ったため小さくなっているが、今後はより大きなシードセットを試す予定である。ゴミ問題や、森林破壊など、環境問題に関するページをシードとして加え、様々な団体の関係などが現れれば非常に興味深い。また、地球環境以外のポータル構築も行う予定である。

参考文献

[1] NASA Goddard Space Flight Center. The earth science portal. http://webserv.gsfc.nasa.gov/ESD/portal.

[2] Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[3] 豊田正史. WWW における関連コミュニティ群の発見. データベースシステム 研究報告 (DBWS2000), No. 122, pp. 307–314, 2000.

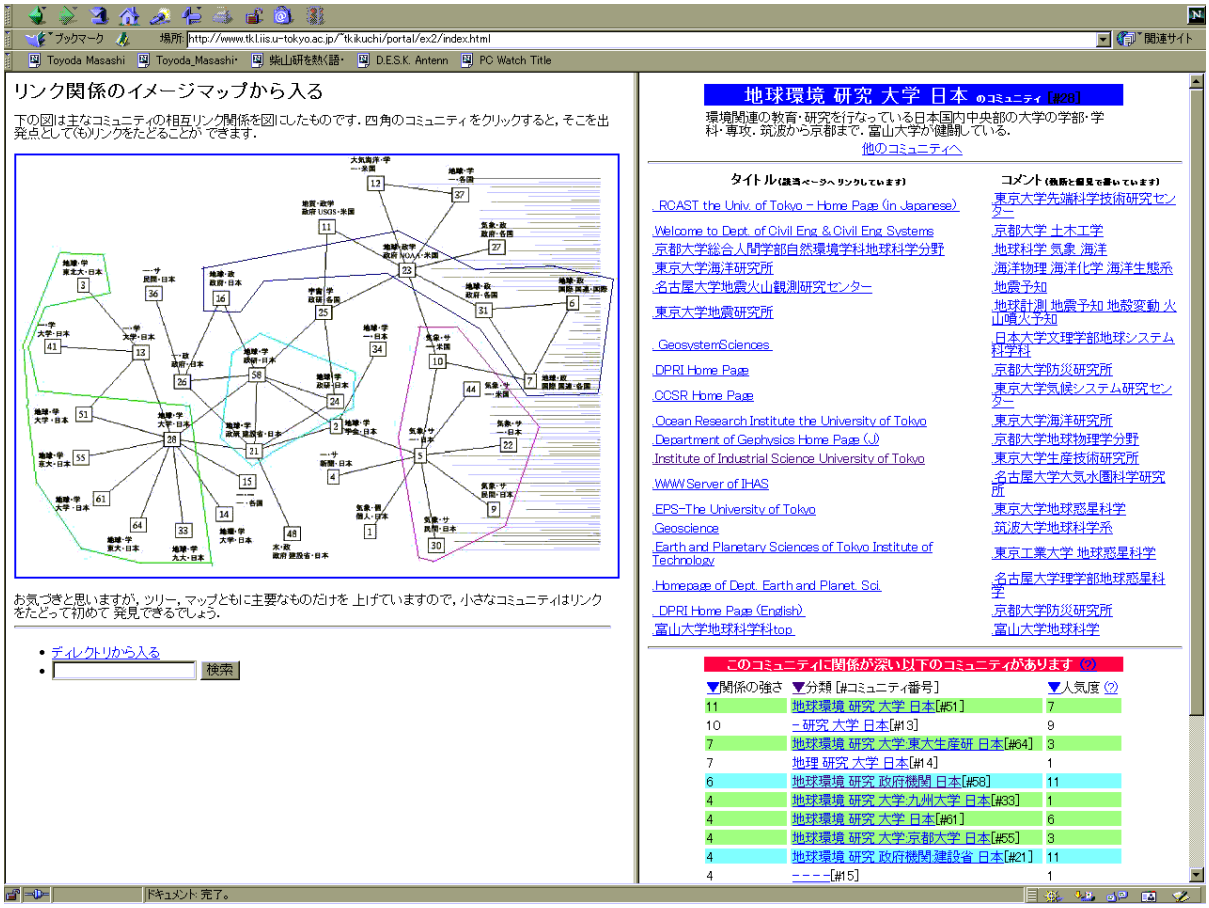
ブックマーク
場所: http://www.tkl.iis.u-tokyo.ac.jp/~tkikuchi/portal/ex2/index.html
関連サイト
Toyoda Masashi
Toyoda_Masashi
柴山研を熱く語
D.E.S.K. Antenn
PC Watch Title
リンク関係のイメージマップから入る
下の図は主なコミュニティの相互リンク関係を図にしたものです．四角のコミュニティをクリックすると，そこを出発点として(も)リンクをたどることができます．
お気づきと思いますが，ツリー，マップともに主要なものだけを 上げていますので，小さなコミュニティはリンクをたどって初めて 発見できるでしょう．
ディレクトリから入る
検索
地球環境 研究 大学 日本 のコミュニティ [#28]
環境関連の教育・研究を行なっている日本国内中央部の大学の学部・学科・専攻．筑波から京都まで．富山大学が健闘している．
他のコミュニティへ
タイトル(該当ページへリンクしています)
コメント(他所と個人で書いています)
RCAST the Univ. of Tokyo - Home Page (in Japanese)
東京大学先端科学技術研究センター
Welcome to Dept. of Civil Eng. & Civil Eng Systems
京都大学 土木工学
京都大学総合人間学部自然環境学科地球科学分野
地球科学 気象 海洋
東京大学海洋研究所
海洋物理 海洋化学 海洋生態系
名古屋大学地震火山観測研究センター
地震予知
東京大学地震研究所
地球計測 地震予知 地殻変動 火山噴火予知
GeosystemSciences
日本大学文理学部地球システム科学科
DPRI Home Page
京都大学防災研究所
CCSR Home Page
東京大学気候システム研究センター
Ocean Research Institute the University of Tokyo
東京大学海洋研究所
Department of Gephysics Home Page (J)
京都大学地球物理学分野
Institute of Industrial Science University of Tokyo
東京大学生産技術研究所
WWW Server of IHAS
名古屋大学大気水圏科学研究所
EPS-The University of Tokyo
東京大学地球惑星科学
Geoscience
筑波大学地球科学系
Earth and Planetary Sciences of Tokyo Institute of Technology
東京工業大学 地球惑星科学
Homepage of Dept. Earth and Planet. Sci.
名古屋大学理学部地球惑星科学
DPRI Home Page (English)
京都大学防災研究所
富山大学地球科学科top
富山大学地球科学
このコミュニティに関係が深い以下のコミュニティがあります (?)
▼関係の強さ
▼分類 [#コミュニティ番号]
▼人気度 (?)
11 地球環境 研究 大学 日本[#51] 7
10 -研究 大学 日本[#13] 9
7 地球環境 研究 大学 東大生産研 日本[#64] 3
7 地理 研究 大学 日本[#14] 1
6 地球環境 研究 政府機関 日本[#58] 11
4 地球環境 研究 大学 九州大学 日本[#33] 1
4 地球環境 研究 大学 日本[#61] 6
4 地球環境 研究 大学 京都大学 日本[#55] 3
4 地球環境 研究 政府機関 建設省 日本[#21] 11
4 ----[#15] 1
ドキュメント: 完了。

図 2: 構築したポータル