最適なウェブサイト構造の調査

株式会社Faber Company （配属先: Faber & Technology）　　B143369 廣中 詩織

ウェブサイト内のページ数が増える → • 重複ページを作成してしまう • サイト構造の把握が難しい →
① 「サイト内ページリストの作成」
② 「リンクネットワークの作成と分析」
③ 「リンクネットワークの可視化」

① サイト内ページリストの作成

目的　重複ページやタイトルの設定ミスなどを発見しやすくする
→ クローラがサイト内を巡回してページをリスト化

問題　ページ（コンテンツ）の取得をいつまで続けるか？
CGIなどによってURLを自動生成するサイトが存在する
→ トップページからリンクをたどる深さを設定 + 時間制限

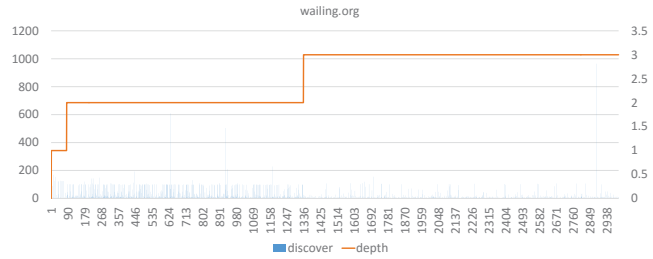


図: クローリング中の新規発見リンク数の推移

表: サイト内のページ一覧（抜粋）。ページごとに設定されているタイトルや見出し、キーワードなどのパラメータを一覧で確認できる。

URL	depth	status_code	content_type	size	title	h1	h2	keywords	description	canonical
.../archives/3853?replytocom=73	3	200	text/html; charset=UTF-8	40148	約7万円 で購入できる3Dプリンタ『3D ...	3Dプリンターのjapanese Makers,約7万円 で購入できる3D...	約7万円 で購入できる3Dプリンタ『3D...	3dプリンタ,メンデル型,激安,格安,3dプリンター	メンデル型の激安3Dプリンターが秋...	.../archives/3853
.../archives/tag/reprap	3	200	text/html; charset=UTF-8	22724	RepRap | 3Dプリンターのjapanese Makers	3Dプリンターのjapanese Makers,RepRap	約7万円 で購入できる3Dプリンタ『3D ...	3Dプリンタ,メンデル型,激安,格安		.../archives/tag/reprap
.../archives/958	1	200	text/html; charset=UTF-8	39994	Arduinoではんだ付け無しで簡単な回路を作る | 3Dプリンターの比...	3Dプリンターのjapanese Makers,はんだ付け無しで簡単...	はんだ付け無しで簡単な回路を作る ～...	ブレッドボード,ジャンパワイヤー,arduino,回路	ブレッドボードとジャンパワイヤーを...	.../archives/958
.../archives/3106?replytocom=35	3	200	text/html; charset=UTF-8	48335	3Dデータ共有サイトより3Dデータをダウンロードして...	3Dプリンターのjapanese ...	3Dデータ共有サイトより3Dデータを...	blade-1,thingiverse,3dプリンター		.../archives/3106
.../archives/1689?replytocom=90	3	200	text/html; charset=UTF-8	44167	家庭用3Dプリンターを見れる場所 | 3D...	3Dプリンターのjapanese Makers,家庭用3Dプリンターを見れる場所	家庭用3Dプリンターを見れる場所～3D...	見学,扱える,体験,展示,デモ,3dプリンター	3Dプリンターを購入検討している...	.../archives/1689

② リンクネットワークの作成と分析

目的　問題のあるコンテンツを発見する
見せたいページ（コンテンツ）がちゃんとリンクされているか？

- 取得したページ間のリンクから**リンクネットワーク**を作成
- リンクネットワークのリンク構造から計算したPageRankとアクセス数から計算したページ訪問率を比較

PageRank
＝ランダムにリンクをたどったときそのページにたどり着く確率

表: PageRankとサイト内部からのページアクセス数とを比較したスコアからリンクの良さをはかることができる可能性がある

page	score
/3d-printer-list	-0.06161
/archives/4105	-0.05593
/archives/2201	-0.05386
/archives/270	-0.03985
/archives/633	-0.03747
/history/サービス	0.01909
/history/動画	0.018975
/history/word	0.018894
/history/管理人のつぶやき	0.018054
/contact	0.017185

良い ↕ 悪い

③ リンクネットワークの可視化

目的　サイトの構造を視覚的に把握する

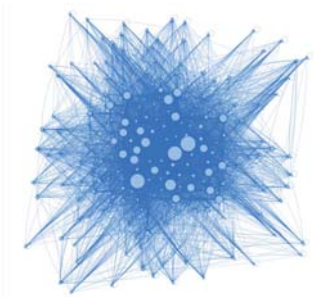
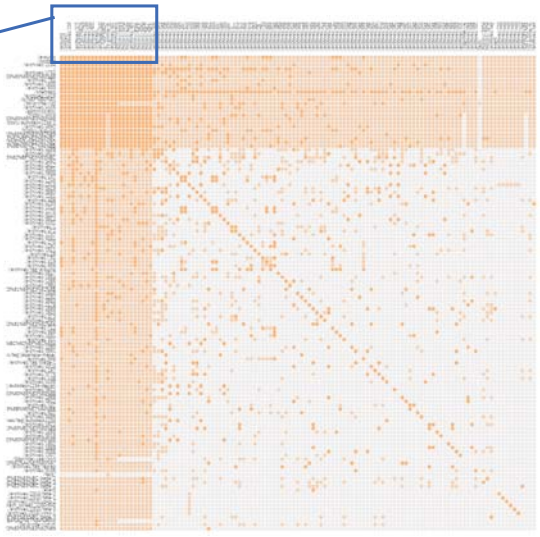
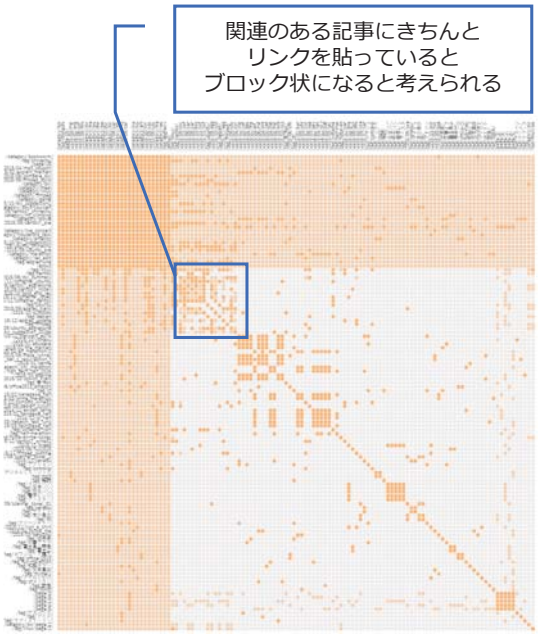


図: コンテンツが乱雑なサイトのネットワーク図（ノードの大きさはPageRankの値をあらわす）
リンクネットワークは密なグラフであるため、ネットワーク図による可視化は適していない。



(a): コンテンツが乱雑なサイト



(b): 構造に少し気を使っているサイト

図: ページ数が同程度のブログサイトから作成したリンクネットワークの隣接行列（PageRankが高い順に並びかえ）
上側と左側の帯は、ページの共通部分にあるリンクである。また、その他の部分をみると、 (a) はランダムにリンクが散らばっているが、(b) はクラスタのようなものができている。