2014/7/4　平成26年度学術情報システム総合WS

学術情報サービスの ログ分析

佐藤翔

同志社大学社会学部教育文化学科（図書館司書課程担当）

自己紹介

- 同志社大学 社会学部教育文化学科 助教
- 学術情報流通論／**利用者行動**分析
- 博士論文：機関リポジトリやCiNii Articlesのログ分析

概要

1. はじめに：ログ分析とは？
2. アクセスログとは？
3. 具体的にどんなことをするの？
4. おわりに：なんとかなるなる！

1. はじめに：

ログ分析とは？

ログとは？

- システムの挙動・運用の記録
  - 保守・メンテナンス等の目的で残される
- Webサービスのアクセスログ
- 図書館の貸出履歴
- データベースの検索ログ
- ILLシステムのログ
- 入退館ゲートの記録 etc…

ログ分析とは？

- システムのログを分析
- 利用者の行動を把握
- システム・サービス運用にフィードバック

ログ分析の短所・長所

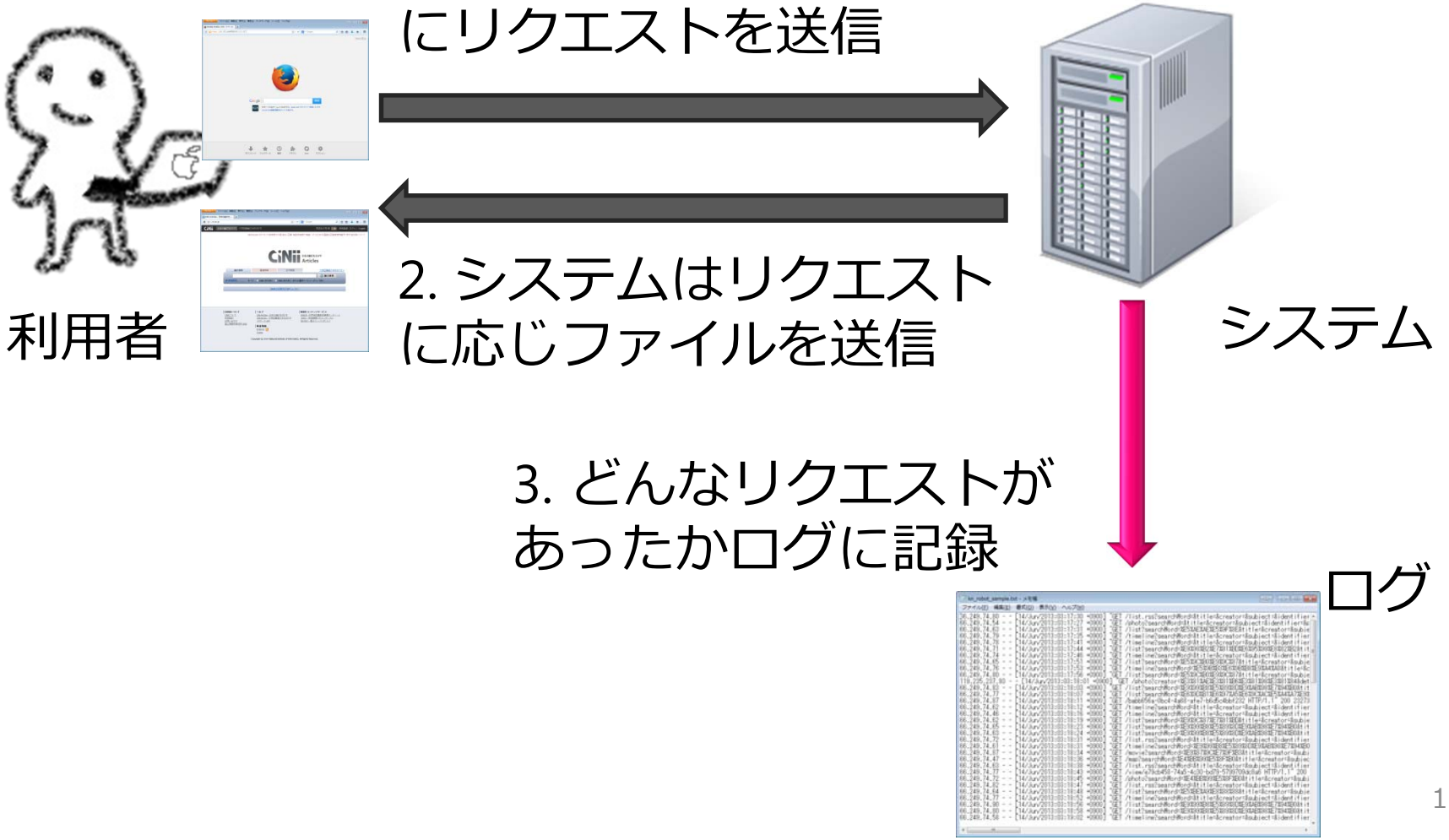
- 短所
  - ログのないものは分析できない←当然！
  - 必ずしも利用分析のためのものではない
  - 利用者の目的・内心は不明
- 長所
  - 全利用の記録である←理論上
  - 新たなデータ収集が不要

ログとは？

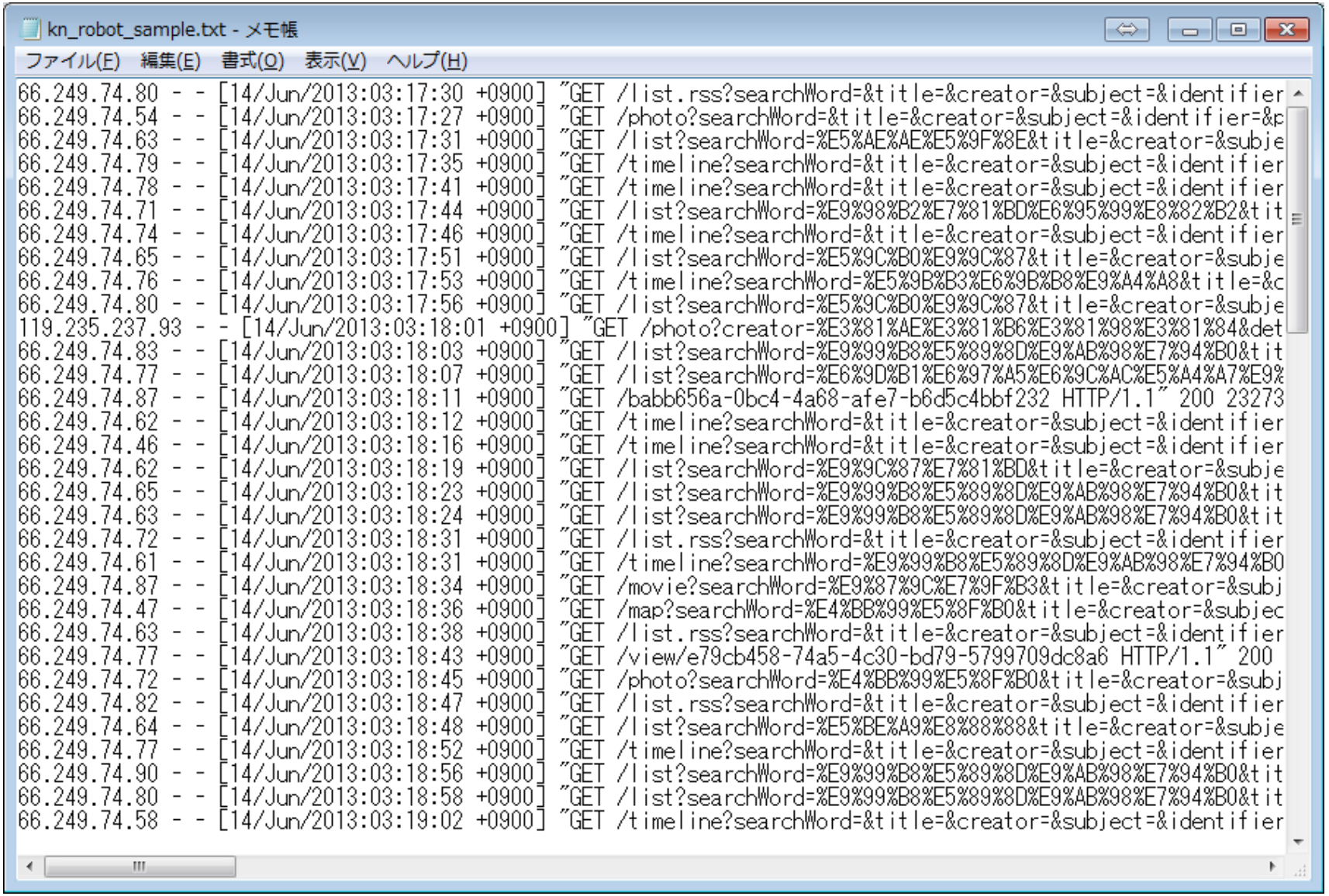
- システムの挙動・運用の記録
  - 保守・メンテナンス等の目的で残される
- **Webサービスのアクセスログ**
- 図書館の貸出履歴
- データベースの検索ログ
- ILLシステムのログ
- 入退館ゲートの記録 etc…

2. アクセスログとは？

アクセスログとは？


1. ブラウザがシステム
にリクエストを送信
2. システムはリクエスト
に応じファイルを送信
利用者
システム
3. どんなリクエストが
あったかログに記録
ログ

アクセスログとは？


kn_robot_sample.txt - メモ帳

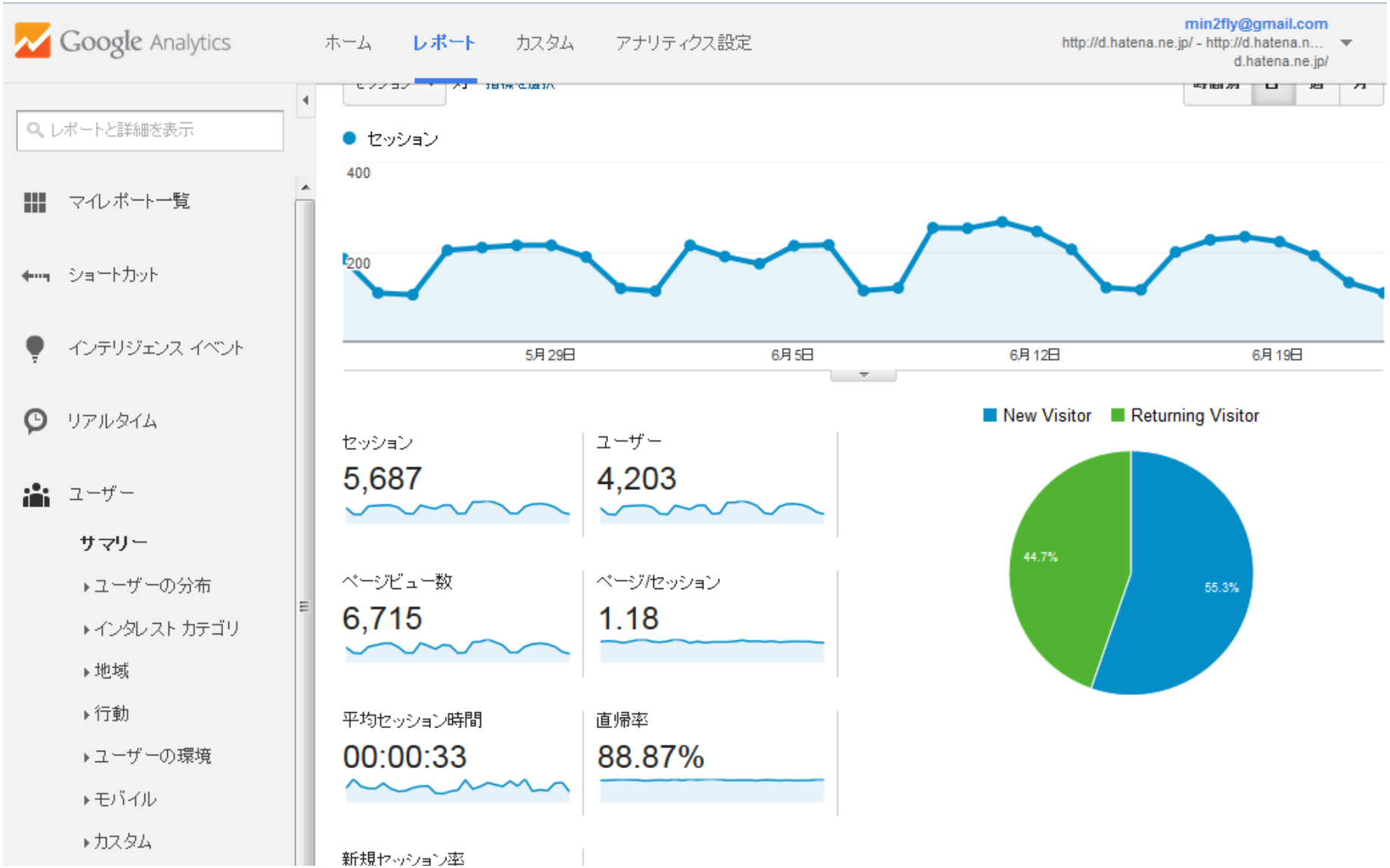
ファイル(F)　編集(E)　書式(O)　表示(V)　ヘルプ(H)

```
66.249.74.80 - - [14/Jun/2013:03:17:30 +0900] "GET /list.rss?searchWord=&title=&creator=&subject=&identifier
66.249.74.54 - - [14/Jun/2013:03:17:27 +0900] "GET /photo?searchWord=&title=&creator=&subject=&identifier=&p
66.249.74.63 - - [14/Jun/2013:03:17:31 +0900] "GET /list?searchWord=%E5%AE%AE%E5%9F%8E&title=&creator=&subje
66.249.74.79 - - [14/Jun/2013:03:17:35 +0900] "GET /timeline?searchWord=&title=&creator=&subject=&identifier
66.249.74.78 - - [14/Jun/2013:03:17:41 +0900] "GET /timeline?searchWord=&title=&creator=&subject=&identifier
66.249.74.71 - - [14/Jun/2013:03:17:44 +0900] "GET /list?searchWord=%E9%98%B2%E7%81%BD%E6%95%99%E8%82%B2&tit
66.249.74.74 - - [14/Jun/2013:03:17:46 +0900] "GET /timeline?searchWord=&title=&creator=&subject=&identifier
66.249.74.65 - - [14/Jun/2013:03:17:51 +0900] "GET /list?searchWord=%E5%9C%B0%E9%9C%87&title=&creator=&subje
66.249.74.76 - - [14/Jun/2013:03:17:53 +0900] "GET /timeline?searchWord=%E5%9B%B3%E6%9B%B8%E9%A4%A8&title=&c
66.249.74.80 - - [14/Jun/2013:03:17:56 +0900] "GET /list?searchWord=%E5%9C%B0%E9%9C%87&title=&creator=&subje
119.235.237.93 - - [14/Jun/2013:03:18:01 +0900] "GET /photo?creator=%E3%81%AE%E3%81%B6%E3%81%98%E3%81%84&det
66.249.74.83 - - [14/Jun/2013:03:18:03 +0900] "GET /list?searchWord=%E9%99%B8%E5%89%8D%E9%AB%98%E7%94%B0&tit
66.249.74.77 - - [14/Jun/2013:03:18:07 +0900] "GET /list?searchWord=%E6%9D%B1%E6%97%A5%E6%9C%AC%E5%A4%A7%E9%
66.249.74.87 - - [14/Jun/2013:03:18:11 +0900] "GET /babb656a-0bc4-4a68-afe7-b6d5c4bbf232 HTTP/1.1" 200 23273
66.249.74.62 - - [14/Jun/2013:03:18:12 +0900] "GET /timeline?searchWord=&title=&creator=&subject=&identifier
66.249.74.46 - - [14/Jun/2013:03:18:16 +0900] "GET /timeline?searchWord=&title=&creator=&subject=&identifier
66.249.74.62 - - [14/Jun/2013:03:18:19 +0900] "GET /list?searchWord=%E9%9C%87%E7%81%BD&title=&creator=&subje
66.249.74.65 - - [14/Jun/2013:03:18:23 +0900] "GET /list?searchWord=%E9%99%B8%E5%89%8D%E9%AB%98%E7%94%B0&tit
66.249.74.63 - - [14/Jun/2013:03:18:24 +0900] "GET /list?searchWord=%E9%99%B8%E5%89%8D%E9%AB%98%E7%94%B0&tit
66.249.74.72 - - [14/Jun/2013:03:18:31 +0900] "GET /list.rss?searchWord=&title=&creator=&subject=&identifier
66.249.74.61 - - [14/Jun/2013:03:18:31 +0900] "GET /timeline?searchWord=%E9%99%B8%E5%89%8D%E9%AB%98%E7%94%B0
66.249.74.87 - - [14/Jun/2013:03:18:34 +0900] "GET /movie?searchWord=%E9%87%9C%E7%9F%B3&title=&creator=&subj
66.249.74.47 - - [14/Jun/2013:03:18:36 +0900] "GET /map?searchWord=%E4%BB%99%E5%8F%B0&title=&creator=&subjec
66.249.74.63 - - [14/Jun/2013:03:18:38 +0900] "GET /list.rss?searchWord=&title=&creator=&subject=&identifier
66.249.74.77 - - [14/Jun/2013:03:18:43 +0900] "GET /view/e79cb458-74a5-4c30-bd79-5799709dc8a6 HTTP/1.1" 200
66.249.74.72 - - [14/Jun/2013:03:18:45 +0900] "GET /photo?searchWord=%E4%BB%99%E5%8F%B0&title=&creator=&subj
66.249.74.82 - - [14/Jun/2013:03:18:47 +0900] "GET /list.rss?searchWord=&title=&creator=&subject=&identifier
66.249.74.64 - - [14/Jun/2013:03:18:48 +0900] "GET /list?searchWord=%E5%BE%A9%E8%88%88&title=&creator=&subje
66.249.74.77 - - [14/Jun/2013:03:18:52 +0900] "GET /timeline?searchWord=&title=&creator=&subject=&identifier
66.249.74.90 - - [14/Jun/2013:03:18:56 +0900] "GET /list?searchWord=%E9%99%B8%E5%89%8D%E9%AB%98%E7%94%B0&tit
66.249.74.80 - - [14/Jun/2013:03:18:58 +0900] "GET /list?searchWord=%E9%99%B8%E5%89%8D%E9%AB%98%E7%94%B0&tit
66.249.74.58 - - [14/Jun/2013:03:19:02 +0900] "GET /timeline?searchWord=&title=&creator=&subject=&identifier
```

2種類のログ解析

1. 生ログ型（ログ型）
   - 自前のサーバに残るログを分析
2. ビーコン型（タグ型）
   - Google Analytics等の方式

Google Analytics



3. 具体的に
どんなことをするの？

アクセスログの見方

133.51.6.255 - - [01/Apr/2009:11:57:18 +0900] "GET /lognavi?name=nels&lang=jp&type=pdf&id=ART0008389255 HTTP/1.1" 302 245 3763 "http://ci.nii.ac.jp/naid/110006390711" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; GTB5; .NET CLR 1.1.4322; InfoPath.1; .NET CLR 2.0.50727; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)"

IPアドレス

133.51.6.255 - - [01/Apr/2009:11:57:18 +0900] "GET /lognavi?name=nels&lang=jp&type=pdf&id=ART0008389255 HTTP/1.1" 302 245 3763 "http://ci.nii.ac.jp/naid/110006390711" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; GTB5; .NET CLR 1.1.4322; InfoPath.1; .NET CLR 2.0.50727; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)"

- リクエスト元の通信機器の番号
- リクエスト元の所属を特定できる
- 個人は特定できない／他制約多し

アクセス日時

133.51.6.255 - - **[01/Apr/2009:11:57:18 +0900]** "GET /lognavi?name=nels&lang=jp&type=pdf&id=ART0008389255 HTTP/1.1" 302 245 3763 "http://ci.nii.ac.jp/naid/110006390711" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; GTB5; .NET CLR 1.1.4322; InfoPath.1; .NET CLR 2.0.50727; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)"

- アクセスのあった日時の記録
- アクセスの集中する日時や、1回の利用時間を特定
- 短時間で極端にアクセスが多いのは・・・

アクセス先ファイル

133.51.6.255 - - [01/Apr/2009:11:57:18 +0900] "GET **/lognavi?name=nels&lang=jp&type=pdf&id=ART0008389255** HTTP/1.1" 302 245 3763 "http://ci.nii.ac.jp/naid/110006390711" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; GTB5; .NET CLR 1.1.4322; InfoPath.1; .NET CLR 2.0.50727; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)"

- アクセスのあったファイルの記録
- どのページが見られていたかがわかる
- 検索に使った言葉等もわかる

参照元（リファラ）

133.51.6.255 - - [01/Apr/2009:11:57:18 +0900] "GET /lognavi?name=nels&lang=jp&type=pdf&id=ART0008389255 HTTP/1.1" 302 245 3763 "**http://ci.nii.ac.jp/naid/110006390711**" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; GTB5; .NET CLR 1.1.4322; InfoPath.1; .NET CLR 2.0.50727; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)"

- どのページのリンクをたどって、そのファイルをリクエストしたかの記録
- アクセス方法／その時の検索語がわかる

User Agent（UA）

133.51.6.255 - - [01/Apr/2009:11:57:18 +0900] "GET /lognavi?name=nels&lang=jp&type=pdf&id=ART0008389255 HTTP/1.1" 302 245 3763 "http://ci.nii.ac.jp/naid/110006390711" "**Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; GTB5; .NET CLR 1.1.4322; InfoPath.1; .NET CLR 2.0.50727; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)**"

- ユーザの環境（ブラウザ、OS、言語等）
- 使っているブラウザやPC/携帯かがわかる
- クローラの特定にも使う（後述）

ログ分析でよく使う項目

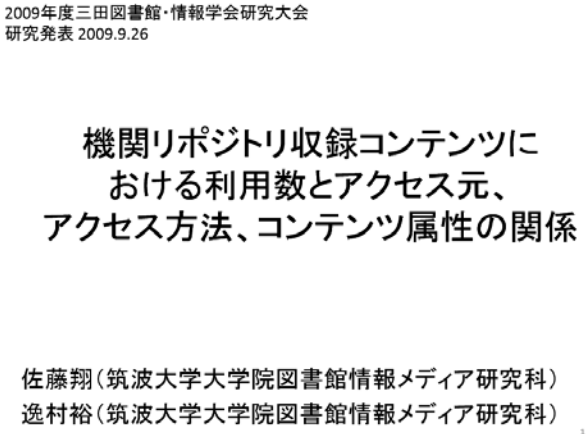
- IPアドレス（利用者の情報）
- アクセス日時
- アクセス先ファイル
- 参照元（アクセス方法）
- User Agent（利用者の環境）

例1：機関リポジトリの分析

- アクセス先ファイル（論文ごと）
  - × IPアドレス（利用者の情報）
  - × 参照元（アクセス方法）

- **参考：**

  佐藤翔, 逸村裕. "機関リポジトリ収録コンテンツにおける利用数とアクセス元、アクセス方法、コンテンツ属性の関係". 三田図書館・情報学会研究大会発表論文集. 2009. http://hdl.handle.net/2241/104869


2009年度三田図書館・情報学会研究大会
研究発表 2009.9.26

機関リポジトリ収録コンテンツに
おける利用数とアクセス元、
アクセス方法、コンテンツ属性の関係

佐藤翔（筑波大学大学院図書館情報メディア研究科）
逸村裕（筑波大学大学院図書館情報メディア研究科）

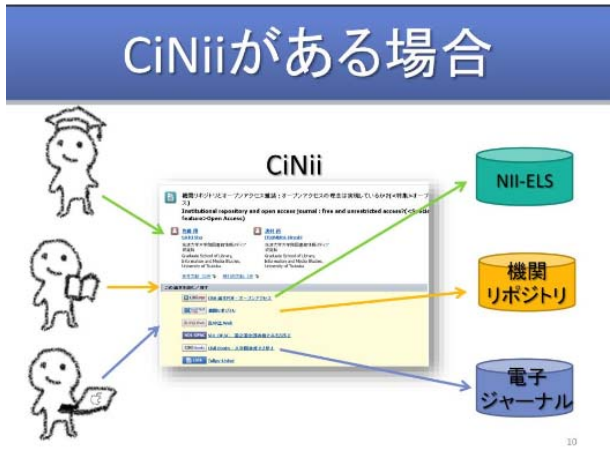
例2：CiNii Articlesの分析

- アクセス先ファイル（論文ごと）

  × 参照元（その前にいたページ）

- **参考：**

  佐藤翔, 大向一輝, 関戸麻衣, 逸村裕. “アクセスログに基づくCiNiiによる本文提供とその利用状況の分析”. 2012年度日本図書館情報学会春季研究集会. 2012. http://hdl.handle.net/2241/117031



例3：NDLサーチの分析

- アクセス先ファイル（検索機能）

  ×　IPアドレス＋日時（利用者の特定）

- **参考：**

  佐藤翔ほか. “アクセスログに基づく国立国会図書館サーチの利用状況の分析”. 第61回日本図書館情報学会研究大会. 2013.
  https://www.dropbox.com/s/je0r0zadfswikvl/JSLIS2013.pptx

実際の手順（１）

- 何を分析したいか／必要なログは何か？
  - ログは量が膨大なので、あらかじめ分析の目的に不要な部分は削除しておく必要がある
- 分析に不必要なログを削除する
  - 不要なファイル、ロボット、外れ値
  - このあたりの作業はプログラムで実行

ロボットって？

- プログラムによる機械的なアクセス
- サーチエンジンのクローラ
  - 検索対象ページを収集するためのもの
- 研究者／技術者等のプログラム
- その他、様々な正体不明のアクセス
- アクセスの大半を占める
- UA等で排除するが・・・

実際の手順（２）

- アクセス元所属が要る場合・・・ドメイン逆引きを実行する
  - IPアドレスから相手のサーバ情報を得る
- 検索語等が要る場合・・・日本語に戻す
  - URLの中では記号になっている
- アクセス方法が要る場合・・・なんらかのURLリストを使って特定する

実際の手順（３）

- その他の分析はケースバイケース
- 自分の場合・・・
  - プログラムを書いて欲しいデータを表にする
  - データベースに落としこむこともある
  - 統計的な分析には市販のソフトも使用
- 詳しくはまた聞いて下さい！

ログ分析の限界

- ロボット問題：ヒトかどうかはわからない
- アクセス元問題：
  - どんなヒトかはっきりしない
  - 同一人物かもはっきりしない
- 内心のことはわからない（UXは不明）
- 正確性には欠ける／全体の傾向は知れる

ログ分析の利点

- 積極的なデータ収集が不要
  - 勝手に残っているからこそのログ
- 全件データが得られる
  - 他は全利用者のデータを得ることは不可能
- 他ではできない詳細分析
  - ログの範囲内ではかなり詳細にできる

4. おわりに：

なんとかなるなる！

ログ分析は難しそう？

- プログラムとか・・・
- データ処理するソフトの使い方とか・・・
- 専門用語が多い・・・
- そのくせ、正確性はないとか言い出す

大丈夫、

なんとかなります

ログ分析は大丈夫

- プログラムとか・・・
  - **簡単な処理でできます**
- データ処理するソフトの使い方とか・・・
- 専門用語が多い・・・
  - **今更ですが、講師は世界史選択のド文系です**
- そのくせ、正確性はないとか言い出す
  - **どの手法も同様です（笑）**

ログ分析は宝の山

- 国内ではやってる人がまだ少ない
  - 片手の数で数えられるくらい
  - 外部発表し放題！
- データは十分すぎるほどある
  - CiNii ArticlesはじめNIIのデータ
  - 皆さんの図書館にもあるかも？

Thank you for your time!