アクセスログを用いた
Webサイト訪問者の行動分析

～Web サイト閲覧者の行動分析による
A社サイト改善の提案～

東京理科大学工学研究科経営工学専攻

修士1年　岩渕隆亮

修士1年　村上尚隆

発表構成

I. 研究背景
II. 関連研究
III. 研究目的
IV. 分析手法
V. 分析結果，考察
VI. 総括，今後の課題

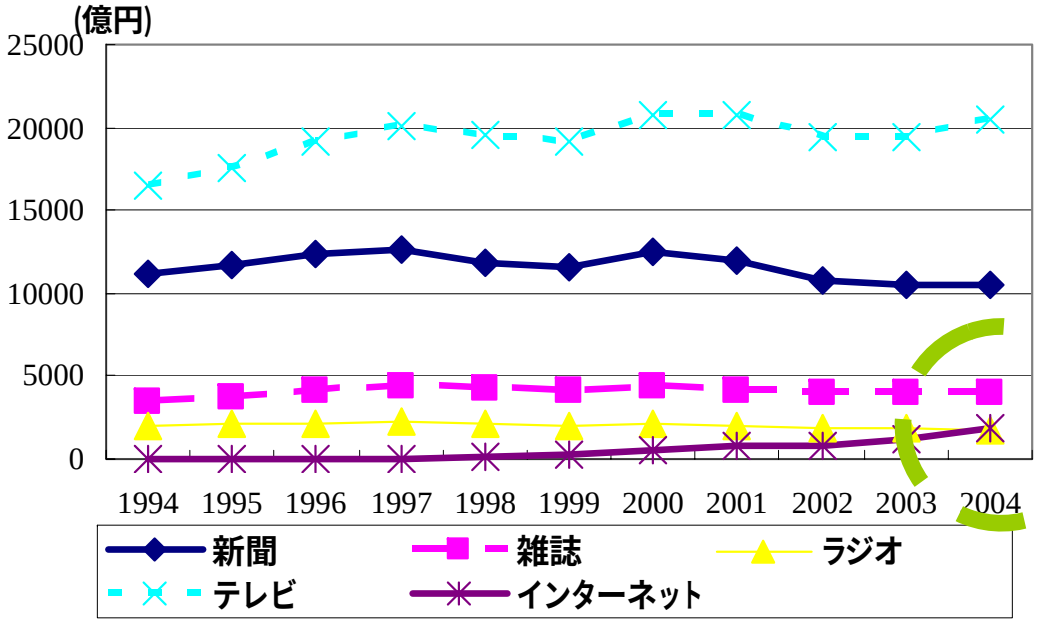
Appendix

1. Visual Mining Studioによる分析
2. S-PLUS,VMSの使用プログラミング

参考文献

研究背景

- 日本におけるインターネットの普及は2005年には4000万世帯を突破すると予測[4]
  - 高速通信設備の普及率の増大
- 2002年以降インターネット広告市場の急拡大



- 2004年インターネット広告費がラジオ広告費を超える（インターネット広告費1814億円，ラジオ広告費1795億円）[5]

図1.媒体ごとの広告費の推移

研究背景

- BtoB-EC（企業間電子商取引），BtoC-EC（消費者向け電子商取引）の市場規模は年々増加 [3]

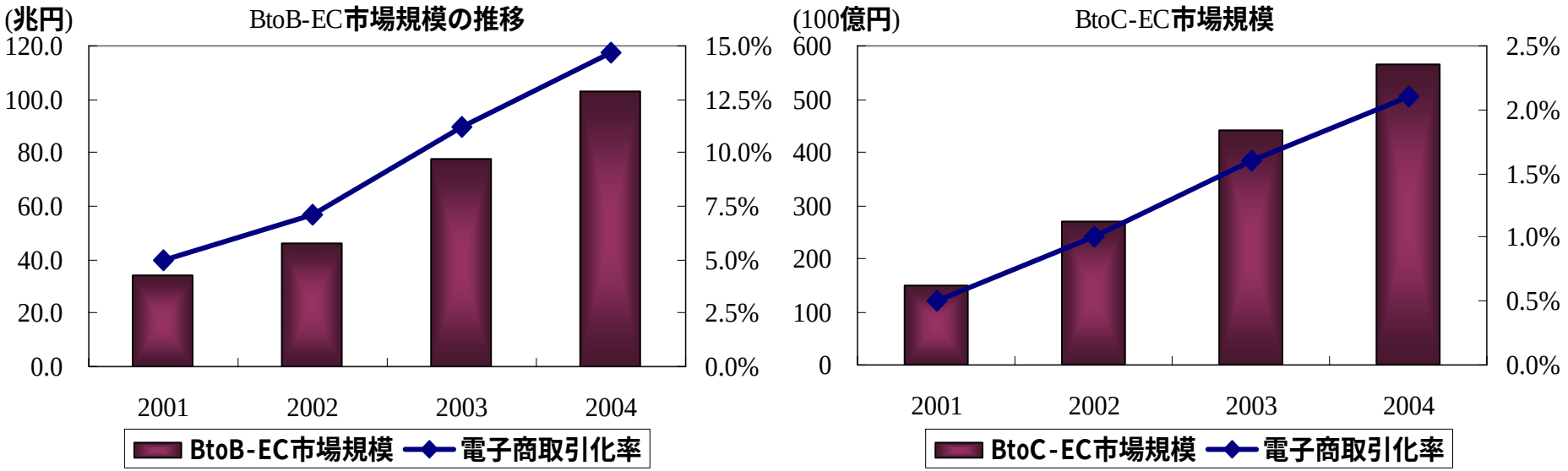


図2.EC市場規模の推移

- Webサイトの訪問者の閲覧行動を分析する必要性が高まる

関連研究

Webサイトの訪問者の閲覧行動を分析する必要性

アクセスログの分析･利用に関する研究が盛んになってきている

- サイト構築者がサイト構造の見直しを図る
- サイトに訪れたユーザの行動支援
  - Web Pageの閲覧者の行動履歴を用いて，閲覧者のクラスタリングを行う[2]（大浦勇亮, 喜連川優，2002）
  - 閲覧者支援を目的とした閲覧者の問い合わせ拡張手法の提案[2]（大浦勇亮, 喜連川優，2002）

研究目的

- **中規模商用サイトのWebアクセスログを用いて，当該サイトの訪問者の閲覧傾向を分類**

分類された各グループに対して効果的なアプローチを提案

- **訪問者の閲覧傾向の分類，特徴づけに関しては主成分分析を用いる**
- **Webアクセスログのフィルタリングは**Visual Mining Studio**を用いる**（Appendix1）

分析手順

閲覧ページ数が2ページ以上の訪問者の
Webアクセスログを抽出

表2を元に分類したWebページと
セッションIDのクロス集計を行う

フィルタリングを行ったアクセスログデータを元に，
当該Webサイトのページを分類

**VMS,S-PLUSを用いて主成分分析を行い
訪問者の閲覧傾向を捉える**(Appendix1)

分析データ

- 本研究では株式会社環の協力の下，
  貴社のWeb Pageのアクセスログを使用した．
  - 閲覧者は，製品情報，製品の導入申し込み，会社概要，
    サービス等の情報を閲覧することができる
- 期間：2005年5月1日から2005年5月31日
- アクセス総数：7060件

表1.アクセスログデータ

項目名	サンプル	項目名	サンプル
年	2005	ユーザエージェント	Mozilla/4.0 (compatible; MSIE 6.0;
月	2	リクエストURL	http://www.nextechcorp.com/
日	27	リファラURL	http://search.yahoo.co.jp/bin/search?p=%A5
曜日	Sun	ユーザID	QiE0ED3T720AAGlH-Ww
時	11	セッションID	QiE0ED3T720AAGlH-Ww
分	44	UNIX時間	1109472272
秒	32	ユニークID	QiE0ED3T720AAGlH-Ww
IPアドレス	61.215.64.10	ディスプレイ縦	800
ポート	61330	ディスプレイ横	600
ホスト名	bh10.ade.point.ne.jp	色深度	32
		訪問回数	1

Webアクセスログのフィルタリング

- Webアクセスログは膨大な量存在する
- 意義あるものを得るためには，データマイニングを行う必要がある

閲覧ページ数が2ページ以上の訪問者のWebアクセスログを抽出

削除したデータ

- 閲覧ページが1ページの訪問者のアクセスログ
- 会社関係者のWebアクセスログ

対象Web Pageの分類

フィルタリングを行ったアクセスログデータを基に，当該Webサイトのページを分類

表2.Webページの分類

対象Webページ	Webページ分類名	含まれる情報
トップページ	**H**(Home)	トップページ
製品概要	**P1**(Product1)	製品の特徴
製品の機能	**P2**(Product2)	製品の詳細な説明，機能の紹介
製品申し込みに関する情報	**O1**(Order1)	料金表，申し込みページ
申し込み完了	**O2**(Order2)	申し込み確認ページ（製品，サービス含む）
オプションサービス	**S1**(Service1)	解析レポート，コンサルティング
その他のサービス	**S2**(Service2)	サポート，メールマガジン，よくあるご質問，提携サービス
その他の情報	**I**(Information)	会社概要，プライバシーポリシー，What's New，プレスリリース

- 詳細に分類すると，訪問者の閲覧傾向が多様化してしまう
- 多様化した閲覧傾向では一定の傾向が捉えられない

クロス集計

**表2を基に分類したWebページと
セッションIDのクロス集計を行う**

各セッション間で
何か傾向がある
のではないか?

表3.クロス集計結果 (一部抜粋)

セッション名	H	P1	P2	O1	O2	S1	S2	I
セッション1	3	11	0	2	0	4	1	0
セッション2	2	1	0	4	0	0	0	0
：	：	：	：	：	：	：	：	：
セッション922	5	6	1	6	1	3	2	3

- 各セッションで訪問者はWebサイトのページを何回閲覧しているか確認する
  - セッション1の訪問者はトップページを3回，製品概要を11回，製品の機能を0回閲覧している

主成分分析による閲覧傾向の分類

VMS,S-PLUSを用いて主成分分析を行い訪問者の閲覧傾向を捉える

主成分分析

- データの中の多くの量的変数を特徴ある少数個の総合的変数に集約し，データの類似関係を明確化する方法
- 分散共分散行列を使用
- データはクロス集計から得られた，各ページの閲覧回数のみを使用

累積寄与率

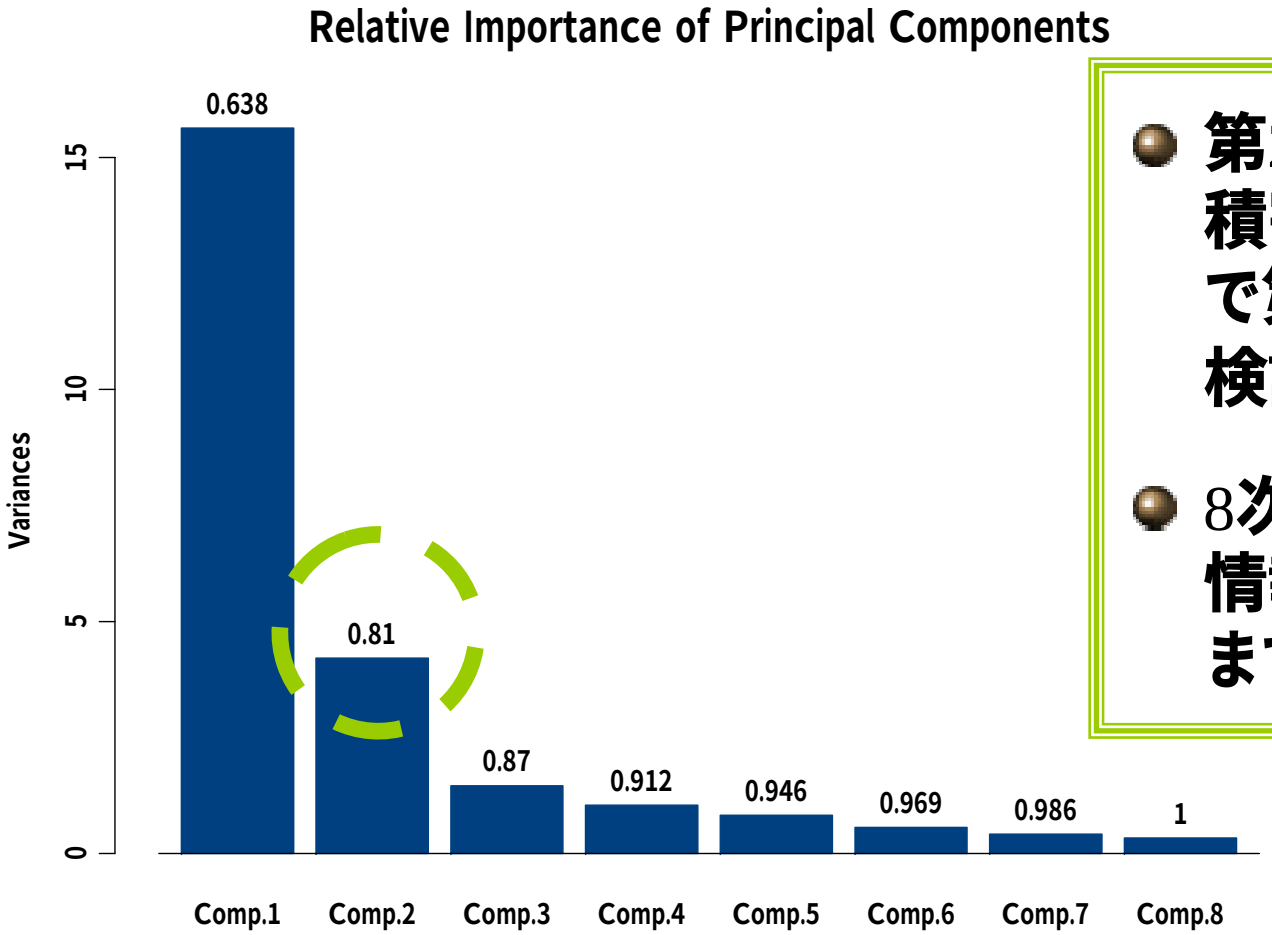


図3.累積寄与率(Comp1:第1主成分)

- 第2主成分までの累積寄与率は0.81なので第2主成分までの検討で十分である
- 8次元のデータを持つ情報のうち第2主成分までで81%説明できる

因子負荷量

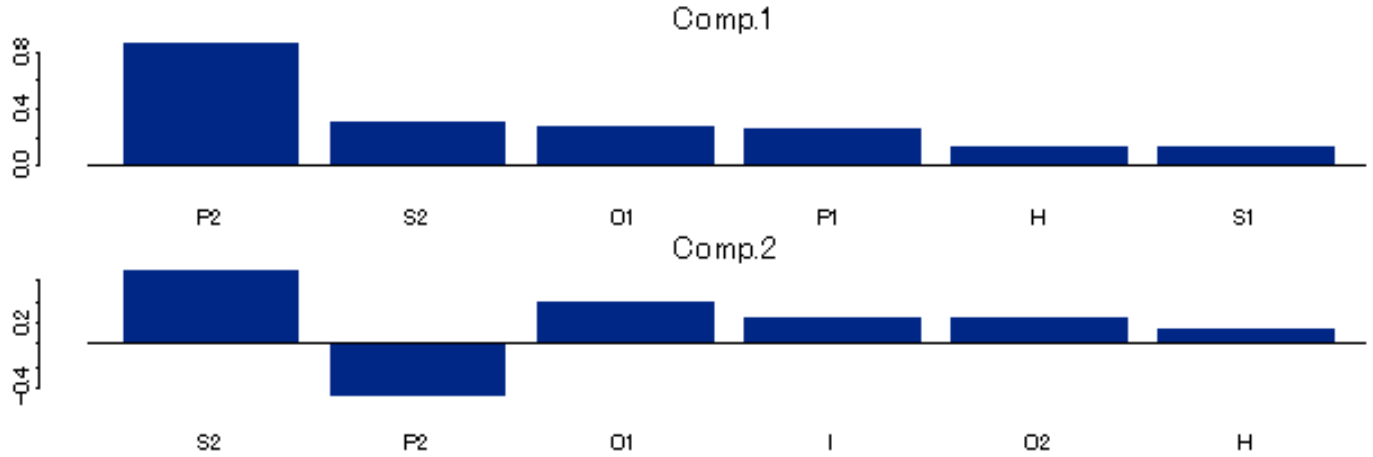


図4.因子負荷量（第1主成分，第2主成分）

第1主成分

- 全体的に値は正，P2の因子負荷量が特に大きい
- Webサイト全体を閲覧し，特に製品の詳細な情報を求めている訪問者

第2主成分

- P2の値が負，それ以外は全て正の値
- S2の値が大きいことから，製品とは関係がないメールマガジン等のその他のサービスに興味がある訪問者

主成分得点の散布図

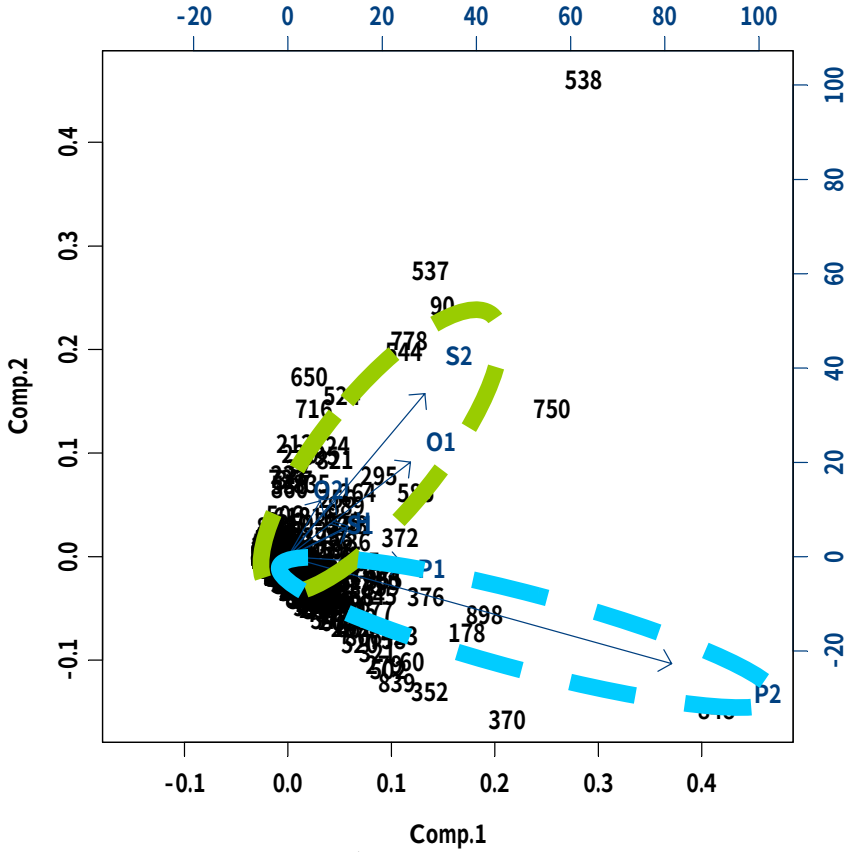


図5.主成分得点の散布図

図5から当該Webサイトにおける訪問者の閲覧傾向として

- P2方向：詳細な製品情報に興味がある訪問者
- S2方向：会社のサービスに興味がある訪問者

2方向に大別できると考えられる

第1主成分に関する考察

- 第1主成分
  - 全体の63.8%がサイト全体を閲覧し，特に製品の詳細情報に興味がある訪問者
  - O1の因子負荷量も正の値であることから，Webサイトの構造が訪問者を申し込みページまでうまく誘導できていると考えられる
    - 各Webページに無料キャンペーンの広告を記載し，申し込みページのリンクを貼っていることが大きな要因ではないか
  - O2の因子負荷量はあまり大きくない
    - 申し込みページまでうまく誘導できているが，申し込みにはつながっていない

申し込みページ：Webサイト内で紹介している製品，サービスを訪問者が購入，契約できるWebページ

表4.第1主成分の因子負荷量

	H	P1	P2	O1	O2	S1	S2	I
第1主成分	0.133	0.254	0.849	0.271	0.072	0.128	0.303	0.104

第2主成分に関する考察

- 第2主成分
  - 全体の17.2%が製品情報よりもその他のサービス（メルマガ，提携サービス）に興味がある訪問者
  - O1の因子負荷量がS2に続いて大きな値である
  - O2の因子負荷量も正の値
    - 申し込み確認ページにたどり着いていることから，製品，サービスの申し込みを済ませている
    - 初めてWebサイトに来て申し込みをするとは考えられにくいので，再訪問である可能性が非常に高い訪問者であると考えられる

表5.第2主成分の因子負荷量

	H	P1	P2	O1	O2	S1	S2	I
第2主成分	0.129	-0.032	-0.455	0.402	0.238	0.119	0.694	0.248

提案アプローチ

- 第1主成分で分類された訪問者に対して
  - 製品情報には興味があるが，申し込みページまでうまく誘導できていない
  - 第1主成分で分類された訪問者はS2（製品の詳細な情報）に興味がある
  - ほぼ全てのページに貼り付けられた，無料キャンペーンの広告を，S2に分類されたページにのみ貼り付け，よりインパクトを持たせることで申し込みページに誘導する
- 第2主成分で分類された訪問者に対して
  - 製品，サービスの情報に対して興味が低い
  - 申し込みページにはうまく誘導できている
  - 少数人数であることもあり新たなアプローチを必要としないと考えられる

総括，今後の課題

総括

- 各ページの閲覧回数のみをデータとして，主成分分析を用いて，訪問者の閲覧行動に関する分析を行った

今後の課題

- 各ページの経路選択の問題に関して，本研究では触れなかった
  - 経路選択により訪問者の行動は大きく変化するかどうか検証
- 細かなページ分類は行わなかった
  - ページ分類を細かに行うことで，コンバージョンに結びつくために，どのようなサイト作りが必要かを詳細に検討できると考えられる
- 第3主成分以下についての閲覧傾向に関しても考察を行うことで，少数グループに対して効果的なマーケティングアクションを提案する

Visual Mining Studioによる分析

図6.Visual Mining Studioによるフロー

1. **データの読み込み**
2. **閲覧ページが1ページのみのセッションを削除**
   - ① **セッションIDが1のものにフラグを立てる**
   - ② **フィルタリング条件の設定**
     table(“ **閲覧ページ数** > 1 → T ”) = = “ T ”

Appendix 1

Visual Mining Studioによる分析

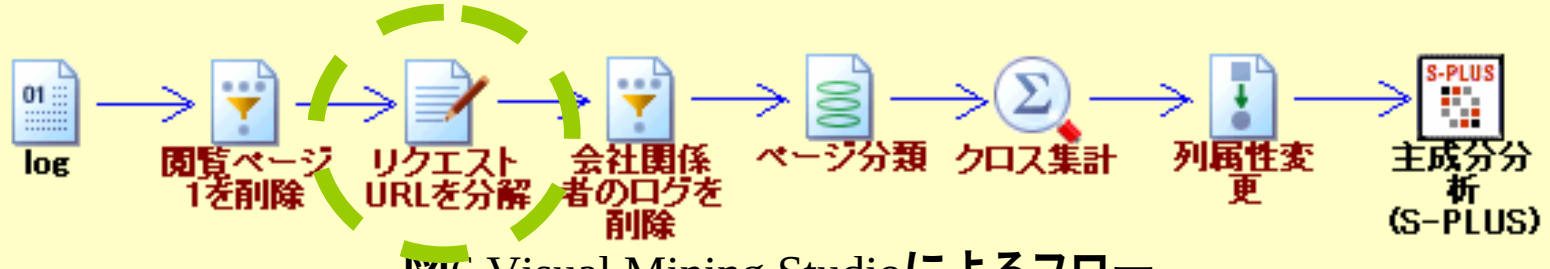


図6.Visual Mining Studioによるフロー

3. **リクエストURLを分解**

分類に使用したURLは以下の通りである（一部掲載）

トップページ：http://www.sibulla.com/index.html

製品概要　：http://www.sibulla.com/info/index.html

http://www.sibulla.com/info/feature.html

製品の機能：http://www.sibulla.com/site/index.html

http://www.sibulla.com/page/index.html

http://www.sibulla.com/etc/index.html

http://www.sibulla.com/seo/index.html

Visual Mining Studioによる分析

リクエストURLの分解手順（例：http://www.sibulla.com/site/index.html）

i. http://www.sibulla.com/を除去
- 全リクエストURLに共通なので，先頭23文字を除去という文字列関数を用いる
- リクエストURLの文字列の長さを数える
  tmp1(“フィルタ”) = strlen(log(“リクエストURL”))
- リクエストURLの24文字目から最後の文字までを取り出す
  tmp2(“リクエストURL2”) = substring(log(“リクエストURL”),24,tmp1(“フィルタ”))
- 除去後はsite/index.htmlとなる

ii. site/index.htmlから製品の機能を表すpageという言葉だけを残す
- /(スラッシュ)を元に，二つの語に分解する文字列関数を用いる
- 上で取り出した文字列を/を境に二つの文字列に分解する
  tmp3(“リクエストURL3”,“リクエストURL4”) = split.str(tmp2(“リクエストURL2”),“/”,“”)
- 元データに列を付け加える
  b = cbind(log,tmp3)

Appendix 1

Visual Mining Studioによる分析

図6.Visual Mining Studioによるフロー

4. **会社関係者のアクセスログを削除**
5. **Webページを分類**
   - **3. で得られた文字列に対して，表2のように分類**
6. **クロス集計**
   - **セッションIDと各分類とのクロス集計**
7. **主成分分析**
   - **6 のクロス集計表を元に主成分分析をS-PLUSで行った**
   - **分析は，分散共分散行列から始めた**

Appendix 2

S-PLUS,VMSの使用プログラミング

S-PLUSの主成分分析のプログラミング

データセット名はbody

```
pr1<- princomp(body)
plot(pr1) #累積寄与率を求める
loadings(pr1)[,1:3]
#第1から第3主成分までの因子負荷量を求める
biplot(pr1)
#主成分得点の散布図を描く
```

参考文献

[1]江尻俊章著，“稼ぐホームページ 損なホームページ”，株式会社アスキー (2004)

[2]大浦勇亮，喜連川優，“Webアクセスログのクラスタリングによる問合わせ拡張支援に関する研究 ”，東京大学生産技術研究所(2002)

[3]経済産業省・ECOM・NTT データ経営研究所 共同，“平成16年度電子商取引に関する実態・市場規模調査”，次世代電子商取引推進協議会(2005)

[4]株式会社情報通信総合研究所 報道発表資料，“http://www.icr.co.jp/info/press/press20020521.html”，(最終閲覧日2005年11月7日)

[5]株式会社 電通 ニュースリリース，“http://www.dentsu.co.jp/news/release/2005/20050060217.html”，(最終閲覧日2005年11月7日)