Webサイト分析・診断技術に関する研究動向

2005年3月2日

NEC インターネットシステム研究所

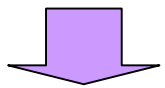
○河合英紀　河野泉　石黒義英　福島俊一

発表の構成

- 1．背景
- 2．Webサイト分析･診断手法
  - ユーザ指向の分析･診断手法
  - サイト指向の分析･診断手法
- 3．リンク整合性チェック
  - 物理的不整合
  - 論理的不整合
- 4．まとめ

1. 背景

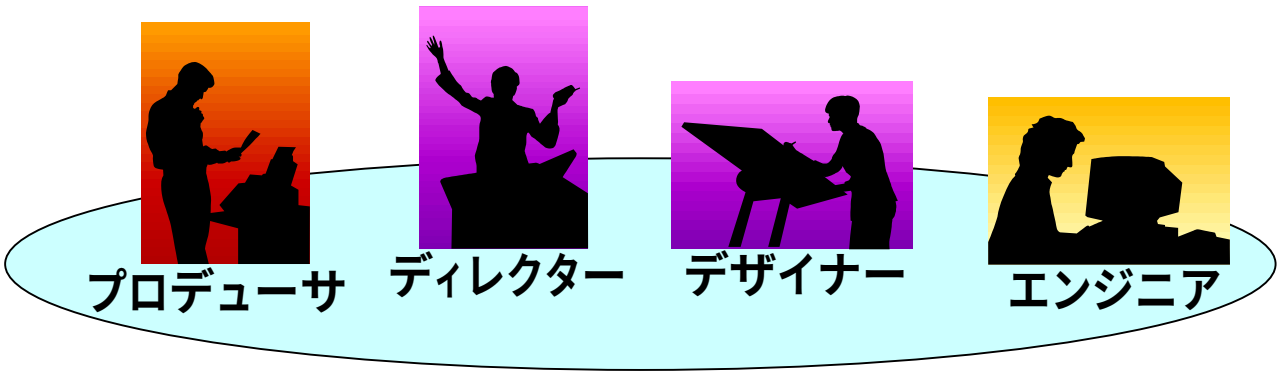
- 「企業の顔」として、Webサイトが普及
  - 大半の企業がWebサイトを開設
    - 上場企業の99.5%（2004年11月 日興アイ・アール株式会社調べ[1]）
    - 中小企業の70.3%（2003年12月 商工中金調べ[2]）
  - 収益に対するWebサイトの貢献度も高い
    - 1位 トヨタ 985億円、2位 マイクロソフト 935億円、3位 ANA 859億円
      - (2004年9月 日本ブランド戦略研究所調べ[3])
- Webサイトの充実に伴い、品質管理が困難に
  - コンテンツの更新頻度向上、サイトの巨大化・複雑化

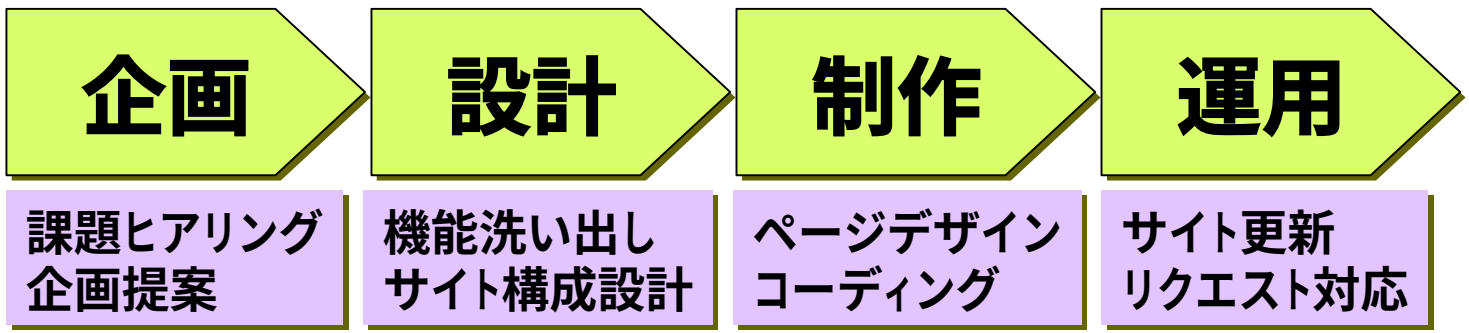


Webサイト分析・診断技術に関するニーズが高まる

1.1 Webサイト構築･運営手法

- 担当者が片手間にコツコツ作成するスタイルから、開発チームが構築･運営するスタイルへ[4-6]
- ホームページ作成ツールから、Webコンテンツマネジメントシステムへ





1.2 Webサイトの診断指標とは？

- 単一の指標で全てを診断できるわけではない
  - 部分的な要素のチェックと、全体的なパフォーマンスの計測
  - 簡便で素早い方法と、精密で時間のかかる方法

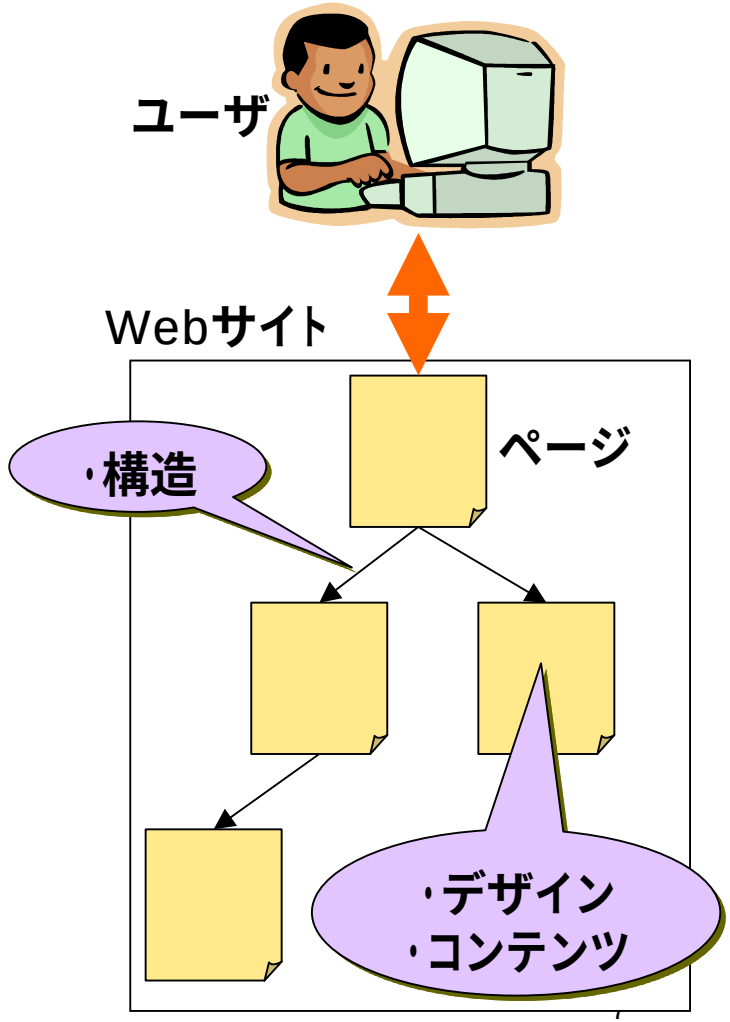
	健康状態	自動車	Webサイト
			
構成要素	・血液/血圧 ・尿 ・視力…	・タイヤ ・ブレーキ ・バッテリ…	・サイト構造 ・コンテンツ ・ページデザイン…
パフォーマンス	・瞬発力 ・持久力 ・生活習慣…	・燃費 ・振動/乗り心地 ・安全性…	・売上/ページビュー ・応答時間 ・セキュリティ…

発表の構成

- 1．背景
- 2．Webサイト分析･診断手法
  - ユーザ指向の分析･診断手法
  - サイト指向の分析･診断手法
- 3．リンク整合性チェック
  - 物理的不整合
  - 論理的不整合
- 4．まとめ

2. Webサイト分析・診断手法

- ユーザ指向の分析・診断手法
  - ユーザの行動をモニタリングして評価
    - ユーザビリティテスト、アクセス/販売ログ解析、視線解析、etc.
- サイト指向の分析・診断手法
  - Webサイト自身をさまざまな特徴量で指標化して評価
  - ユーザの行動や経験を、ユーザモデルやガイドラインとして取り込む手法も存在
    - サイト構造解析、ガイドライン検査、競合サイト比較、リンク整合性チェック、etc.

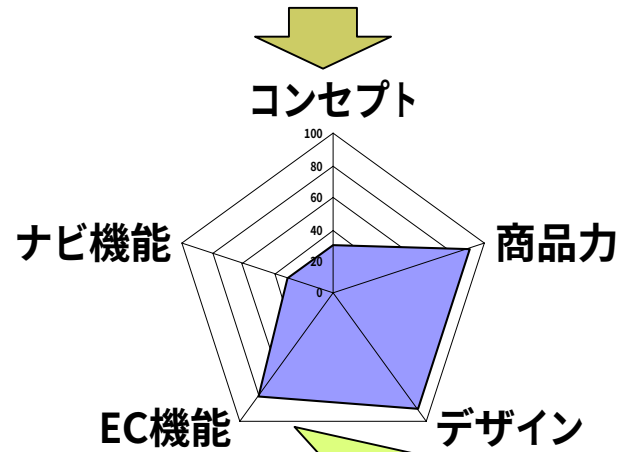


2.1 ユーザ指向の分析・診断手法

2.1.1 ユーザビリティテスト

- 【目的】サイトの使いやすさの診断
- 【主な手法】テストユーザによる評価、専門家によるヒューリスティック評価、アンケートや、グループインタビュー等[7-10]
- 【関連事例】
  - Jacob Nielsenによる、著作・研究報告多数[9-11]
  - 必要十分なテストユーザの数（Magic number 5）にまつわる議論@CHI '03[12]
  - テストで得られた知見をガイドライン化し、チェックを自動化する研究もホット。（→ c.f. 2.2.2 ガイドライン検査）

・サイトのコンセプトは明確か？
・商品の品揃えは豊富か？
・FAQは充実しているか？
・ページタイトルは具体的か？
・リンクは見易いか？
・用語は統一されているか？
・・・・

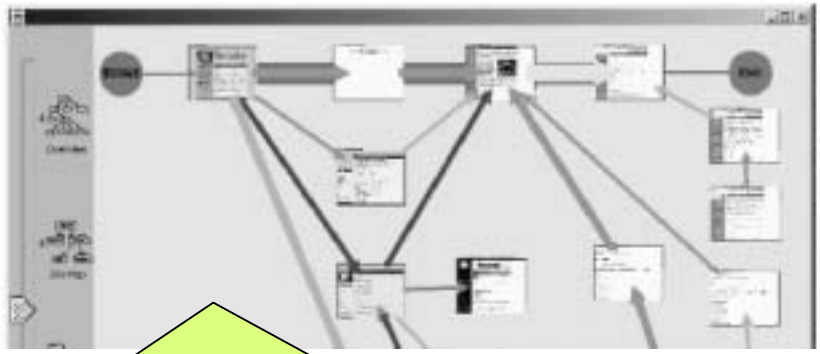


商品は豊富で魅力的だが、ユーザが欲しいものを見つけられていない

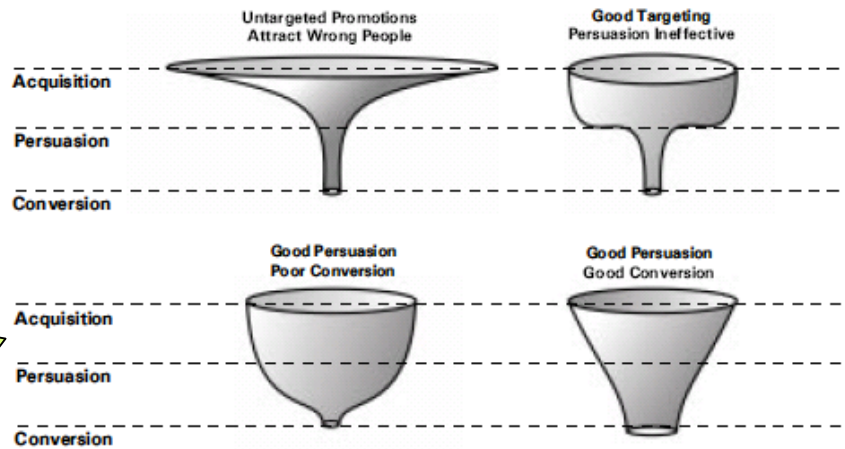
2.1 ユーザ指向の分析・診断手法

2.1.2 アクセス/販売ログ解析

- 【目的】実際の訪問/販売履歴から、サイトの構造や販売力を評価
- 【主な手法】アクセスログによる訪問履歴の解析・可視化と、訪問履歴と売上等の経営指標との関連付け
- 【関連事例】
  - 訪問履歴の解析・可視化
    - WebQuiltプロジェクト@UC Berkeley[13]
  - 経営指標との関連付け
    - E-Metrics@NetGenesis[14]



WebQuilt[13]の訪問経路可視化
線の太さ：ユーザの数
線の色：かかった時間



E-Metrics[14]
Customer Life Cycle Funnel
収集→説得→購買のうち、どこがボトルネックかを可視化

2.1 ユーザ指向の分析･診断手法

2.1.3 視線解析

- 【目的】ページデザイン、コンテンツの評価
- 【主な手法】テストユーザの視線計測と、その履歴を可視化
- 【関連事例】
  - eye-trackingプロジェクト@Cornell大学 [15]

閲覧箇所の分布を可視化

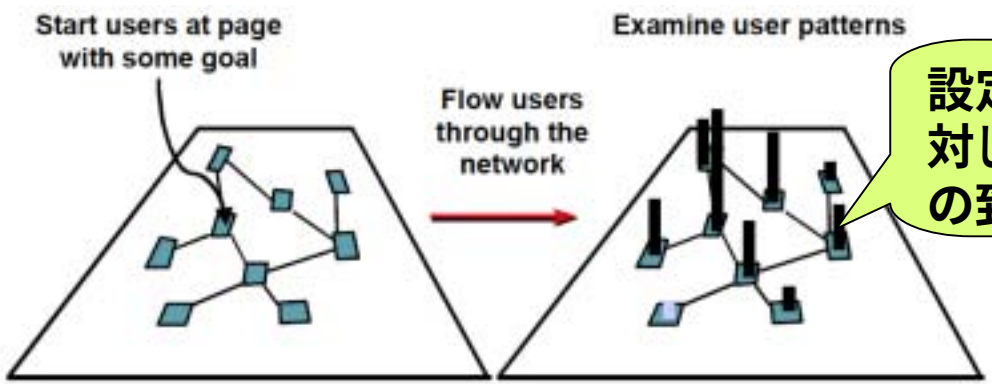
閲覧履歴

2.2 サイト指向の分析・診断手法

2.2.1 サイト構造解析

- 【目的】サイト内でユーザが所望するページへの経路とその見つけやすさを予測
- 【主な手法】キーワードで表現される閲覧目的に対し、アンカー文字列やリンク周辺の文字列との類似度から、各リンクの押され易さを数値化し、ターゲットページへの到達率を計算。
- 【関連事例】
  - InfoScent Bloodhound @PARC[16]

タスクの設定：ゼロックスのサイトで、「ハイエンドのコピー機」を探す

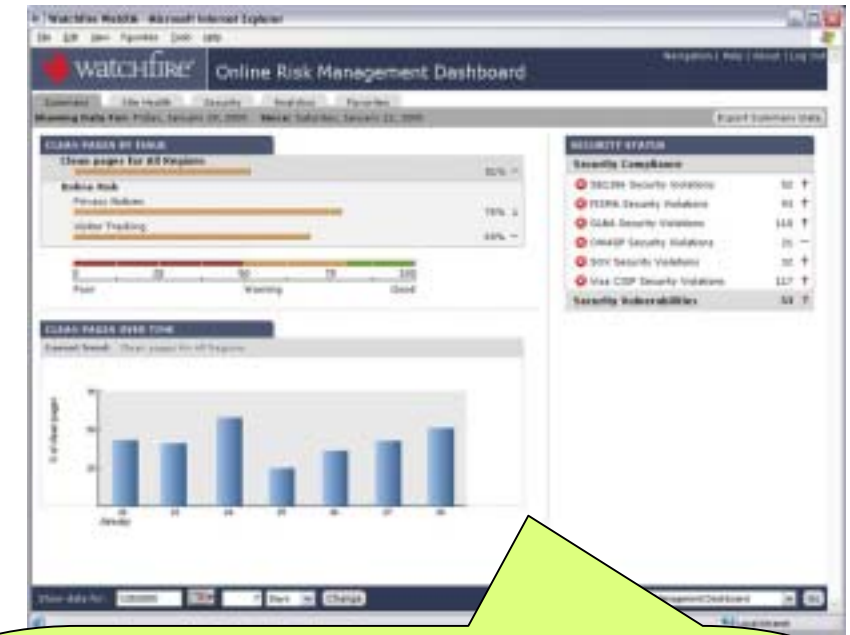


設定されたタスクに対して、各ページへの到達率を計算

2.2 サイト指向の分析・診断手法

2.2.2 ガイドライン検査

- 【目的】人間が定めた基準をどの程度満たしているかを検査
- 【主な手法】ユーザビリティガイドライン、アクセシビリティガイドライン、HTML文法など、定められた基準との一致度をチェックしスコア化。逆に、人間が評価を与えたサイトを入力とし、機械学習を使って判別ルールを導出する手法もある。
- 【関連事例】
  - チェックツール（製品）
    - WebXM@WatchFire[18]
  - 判別ルール導出
    - WebTango@UC Berkeley→ワシントン大[19]



WatchFire WebXM[18]
(1)品質（スペルミス、レスポンスタイム等）
(2)セキュリティ（FORMの安全性等）
(3)プライバシー（P3P対応等）
(4)アクセシビリティ（W3C勧告の準拠度等）

次スライドで説明

2.2 サイト指向の分析・診断手法

2.2.2 ガイドライン検査（続き）

- WebTango@UC Berkeley→ワシントン大[19]
  - サイトの良し悪しをWebデザインに関する様々な文献から収集した、157の指標を使って評価
  - あらかじめ審査員が付与した639サイトの評価結果から、「良い」サイトの条件となるルールを決定木等で学習

テキスト要素	31
リンク要素	6
グラフィック要素	6
テキストフォーマット	24
リンクフォーマット	3
グラフィックフォーマット	7
ページフォーマット	27
ページパフォーマンス	37
サイト構造	16

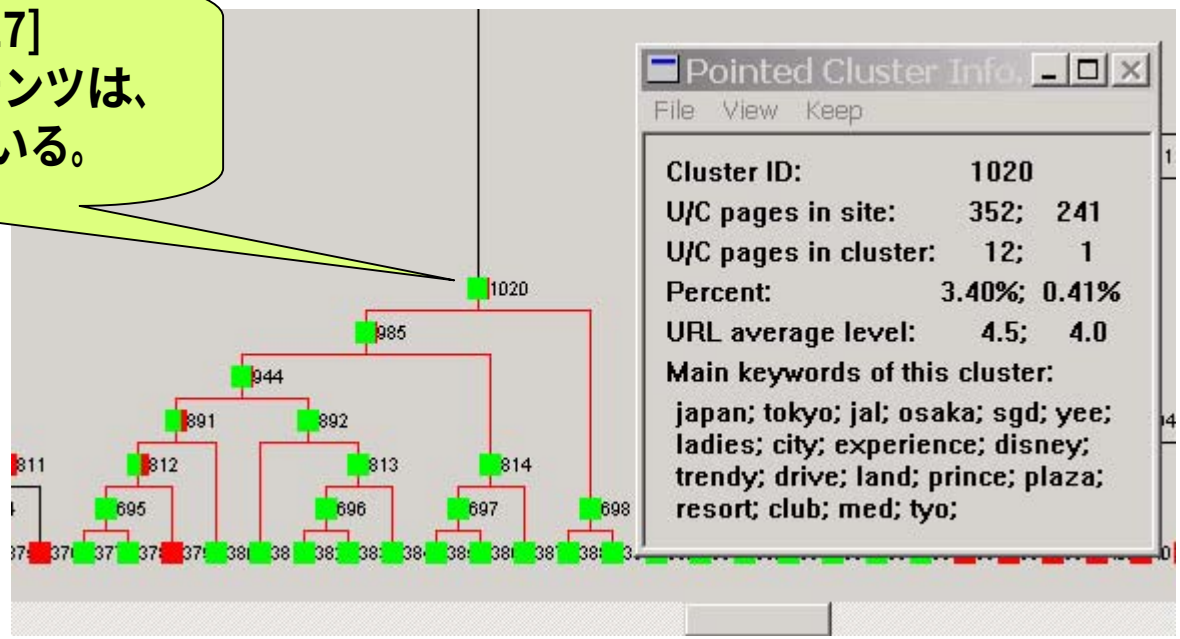
診断結果と根拠の例
Predicted quality: poor
Rationale: if ((Italicized Body Word Count is missing OR (Italicized Body Word Count > 2.5)) AND (Minimun Font Size is missing OR (Minimum Font Size <=9.5)))

2.2 サイト指向の分析・診断手法

2.2.3 競合サイト比較

- 【目的】競合他社の動向調査と差別化
- 【主な手法】自社・他社のWebサイトを単語ベースで階層的にクラスタリング。各クラスタに含まれるそれぞれのサイトのページを色分け
- 【関連事例】
  - Visualizing Web Site Comparisons@National Univ. of Singapore[17]

旅行代理店サイト比較の例[17]
日本への旅行に関するコンテンツは、
緑のサイトに偏って含まれている。



2.2 サイト指向の分析・診断手法

2.2.4 リンク整合性チェック

- 【目的】リンクの整合性に着目した品質保持
- 【主な手法】診断対象のサイトのページを収集してエラーを検出。
- 【関連事例】
  - ガイドライン検査やWebコンテンツマネジメントシステム製品の機能の一つとして提供[18,20]

LinkScan@Elsop Corp. [20]
サイト内で発生したエラー一覧。デッドリンクの他にも、文書タイトルの有無、孤立ページ、応答時間の遅いページ等のチェックが可能

2.3 サイト分析・診断手法まとめ

		構成要素			パフォーマンス	
		サイト構造	コンテンツ	ページデザイン	性能運用状態	経営指標
ユーザ指向	ユーザビリティテスト	○	○	○	○	
	ログ解析	○				○
	視線解析		○	○		
サイト指向	サイト構造分析	○	○			
	ガイドライン検査		○	○		
	競合サイト比較		○			
	リンク整合性	○			○	

※（ご参考）サイト指向の自動診断については、Melody Y. Ivory, et al. "The State of the Art in Automating Usability Evaluation of User Interfaces"[21]に詳しく分類されている。

発表の構成

- 1．背景
- 2．Webサイト分析･診断手法
  - ユーザ指向の分析･診断手法
  - サイト指向の分析･診断手法
- 3．リンク整合性チェック
  - 物理的不整合
  - 論理的不整合
- 4．まとめ

3.リンク整合性チェック

- なぜ、リンク整合性チェックなのか？
  - リンク先の移動や消滅により、経年劣化しやすい
  - 比較的簡便な方法で、サイト全体を検査可能
  - 運用面での問題点が表面化しやすい

定期的な検査が必要かつ可能で、サイト運用における問題点の早期発見と解決につながるのではないか？

Webサイト診断手法における血圧検査

3.1 リンク不整合とは

サイト閲覧者の期待や、サイト製作者の意図に反する効果をもたらすリンク

- **物理的不整合（リンク先へアクセス不可能）**
  - デッドリンク、サーバーエラー等

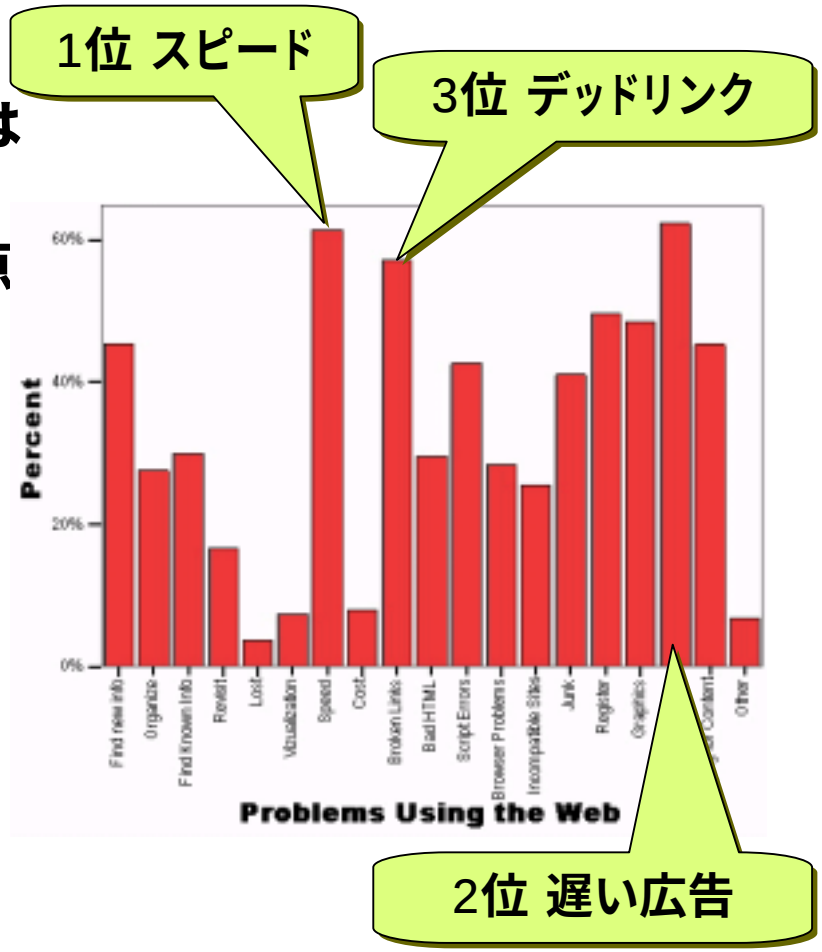
➡ **従来ツール（リンクチェッカー）で検出可能**

- **論理的不整合（アクセス可能だが不適切）**
  - 間違いリンク
  - リンク元表記の不統一
  - 幽霊リンク

➡ **従来ツールで検出不可能～人手によるしらみつぶしチェック**

3.2 物理的不整合の関連研究

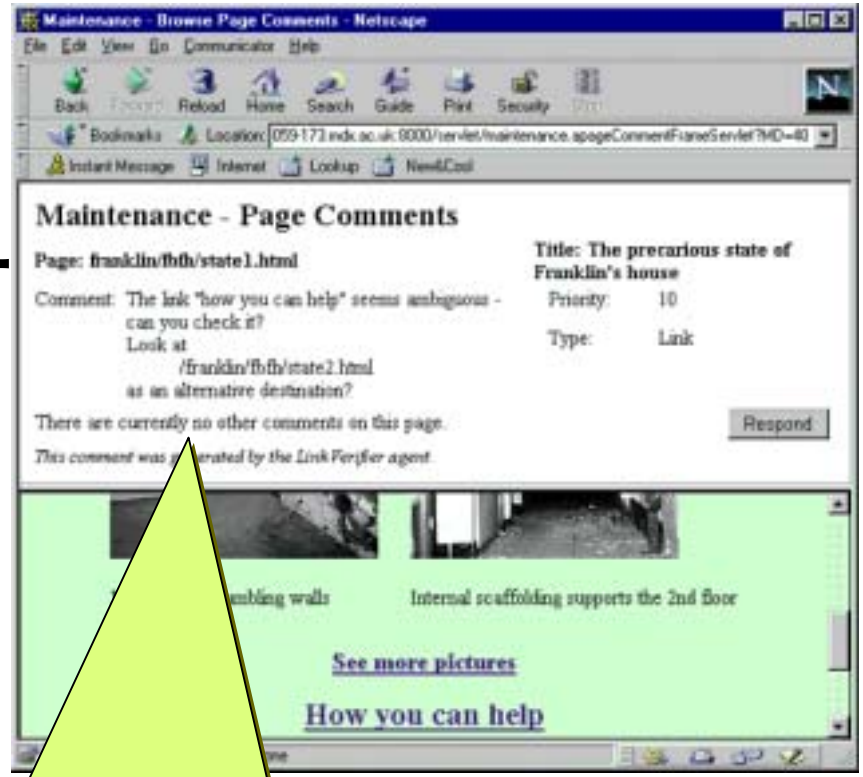
- 黎明期から問題点として重要視
  - HTML文書の平均的なライフサイクルは75日[22](1996年)
  - WWWユーザサーベイで、重要な問題点の第3位に[23](1998年)
  - AOLのサーバーにリクエストされたリンクの5～8%がデッドリンク[24](1998年)
- 自動修正に関する研究多数[30]
  - ページを特定できる文字列を利用
    - Lexical Signatures@Pennsylvania State University, NEC[25]
  - ページのリンク元を利用
    - Link Authority@芝浦工大、図書館情報大、筑波大[26]



3.3 論理的不整合の関連研究(1)

- 物理的不整合に比べ未開拓
  - Meaningful Link Verification @Middlesex大学[27]
    - リンク周辺の情報と、リンク先ページの情報の近さをスコア化
    - リンク先に一致するキーワードがなかったり、逆にどのページにも一致する場合、現在のリンク先よりも適切なリンク先があると見なす
  - サイト品質管理のためのリンク不整合検出@NEC[28,29]

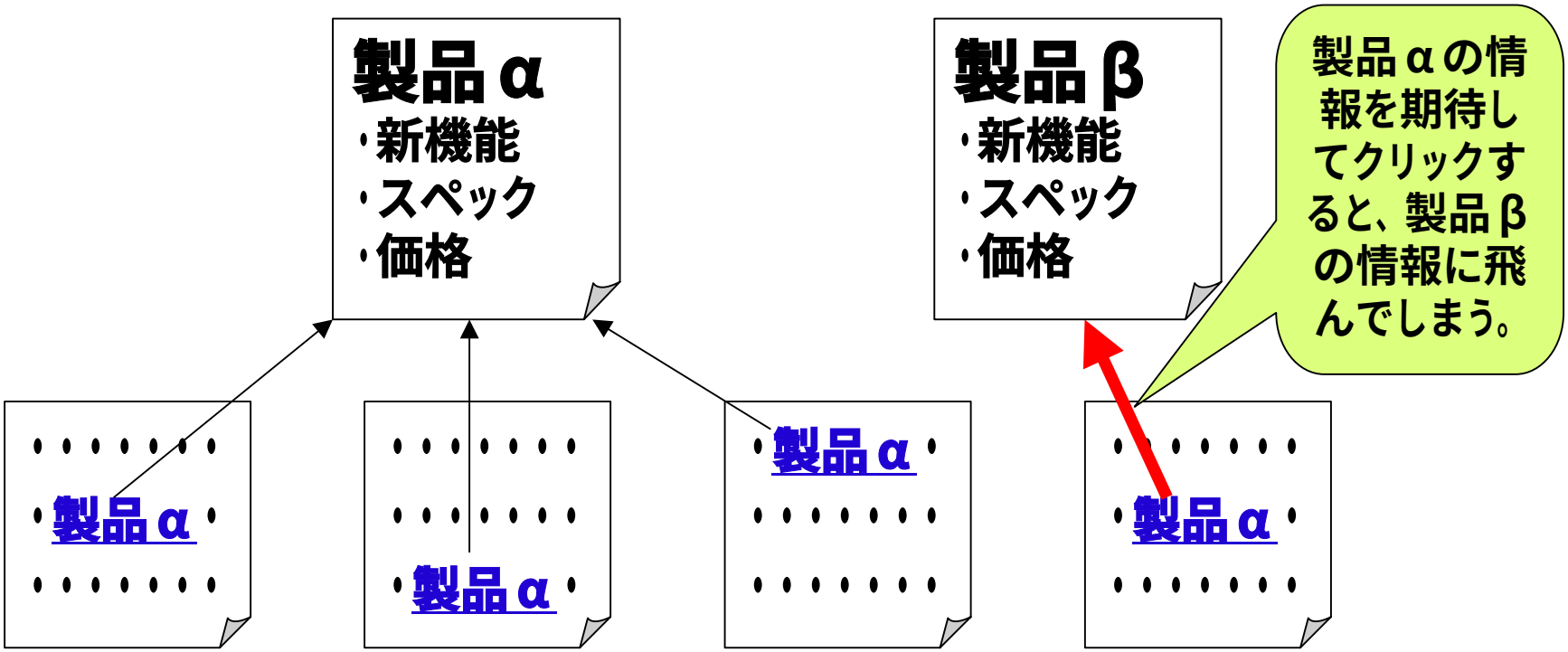
次スライド以降で説明



Meaningful Link Verification[27]
警告:「How you can help」というリンクは曖昧

3.3.1 間違いリンクの例

- アンカー文字列から期待される内容と、リンク先の文書の内容が異なるリンク



・同一アンカー文字列のリンクでグループ化するとリンク先が異なる

3.3.2 論理的不整合候補の検出例

ルール(1) アンカー文字列が同一だがリンク先が異なる場合の例：

リンク数	リンク元 *s*(代表)	リンク先 *t*	アンカー文字列 *a*
2570	URL_3	URL_1 (製品αの情報)	製品α
55	URL_4	URL_2 (製品βの情報)	製品α

○正しい

×間違い

提案方式のメリット：
・正解がどんなリンクか、比較しながらチェックできる
・一度のチェックでまとめて不整合をチェックできる

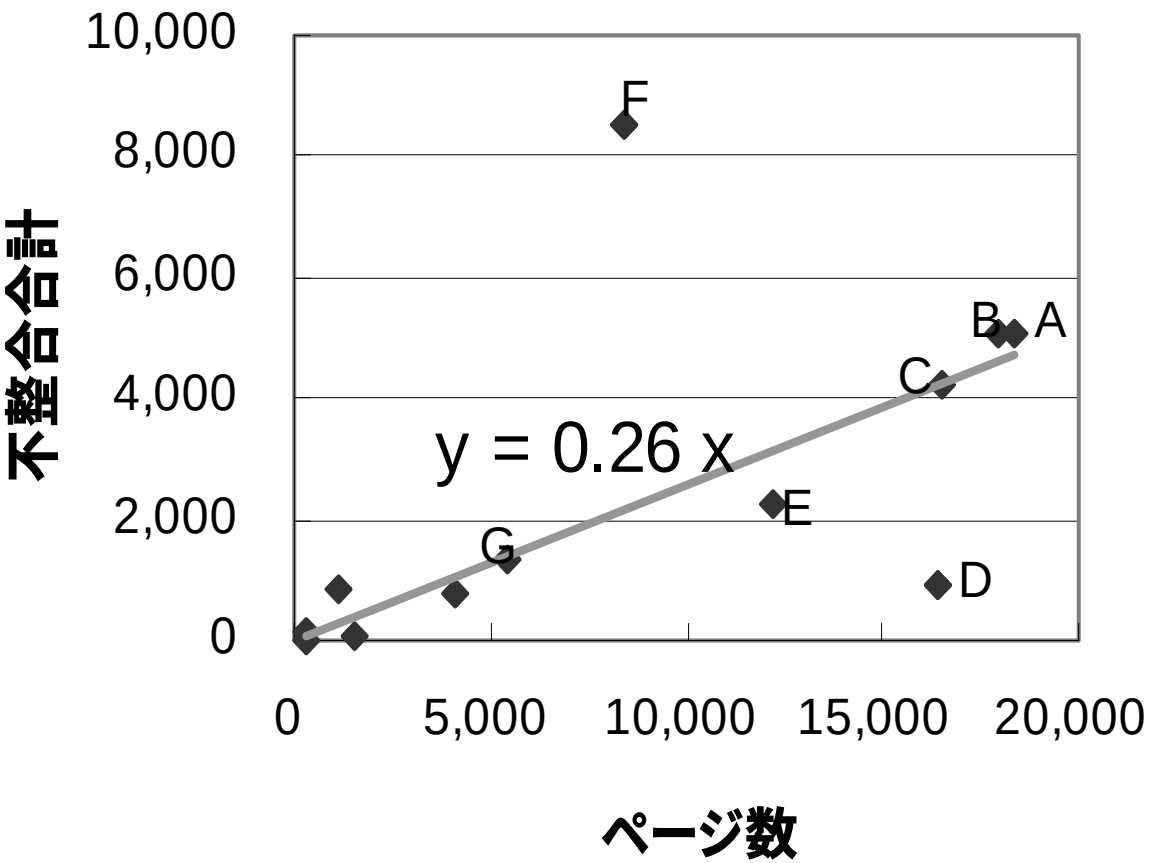
3.3.3 リンク不整合の実態

- 12サイトに合計3万件弱のリンク不整合が存在
- 論理的不整合も物理的不整合と同程度検出された

サイト	ページ数	リンク数	物理的不整合	論理的不整合	合計
A	18,389	1,263,562	2,766	2,287	5,053
B	17,900	372,322	494	4,570	5,064
C	16,560	151,195	1,086	3,135	4,221
D	16,393	219,318	330	556	886
E	12,161	221,430	1,511	711	2,222
F	8,430	191,437	7,236	1,277	8,513
G	5,450	49,141	39	1,276	1,315
H	4,072	52,701	358	430	788
I	1,532	9,977	78	8	86
J	1,130	35,796	0	873	873
K	339	11,798	4	21	25
L	331	7,026	1	110	111
合計			13,903	15,254	29,157

3.3.4 規模とリンク不整合の関係

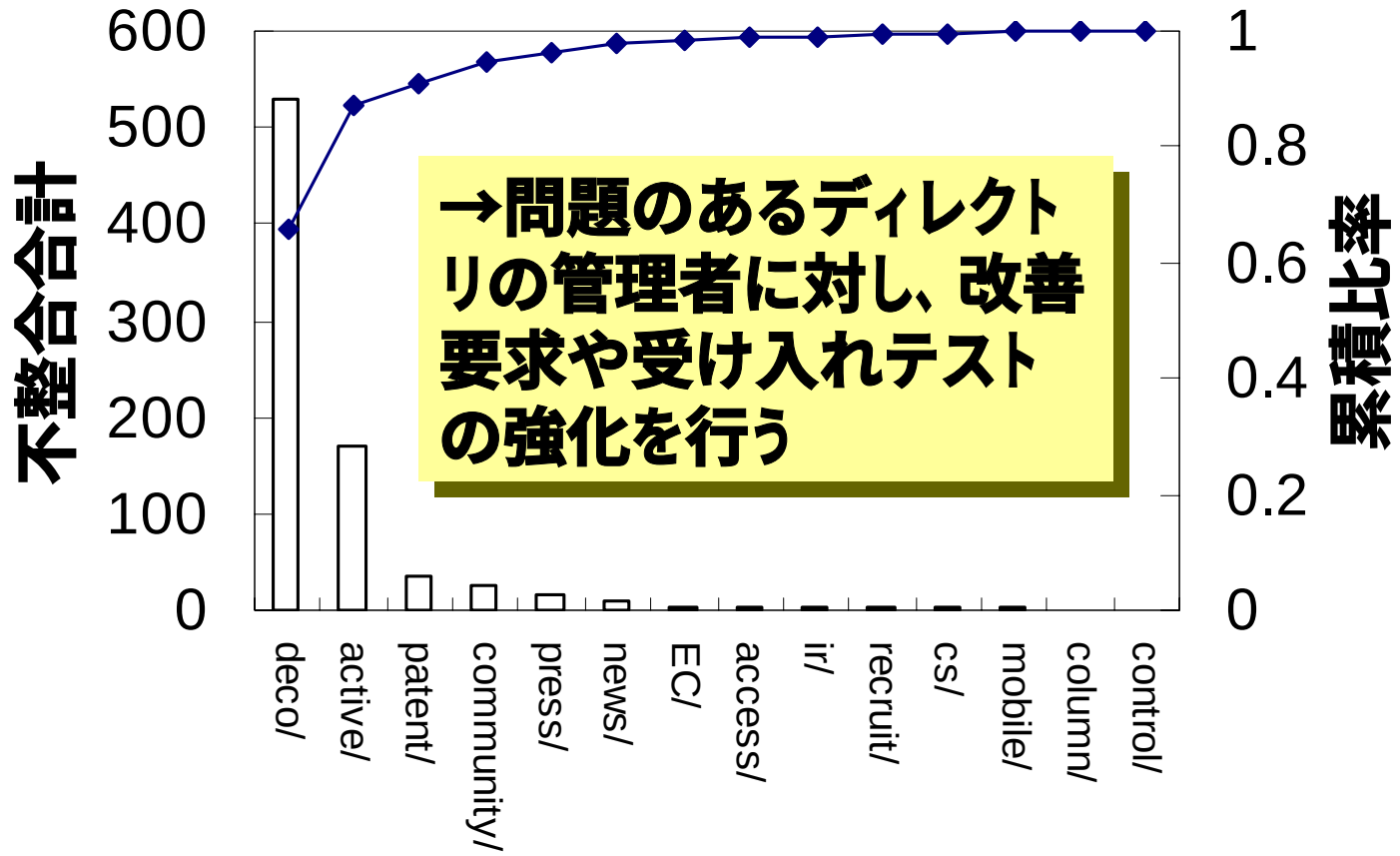
- 平均的なサイトのリンク不整合の発生率は、サイトの種類や規模によらず、ほぼ一定



→「メンテナンスが優れたサイト」における、リンク不整合の発生率は1ページ当たり約0.26件以下

3.3.5 ディレクトリ別分布

- 特定のディレクトリ(　担当部署)下で管理されているコンテンツに不整合が集中→運用上の問題



3.3.6 論理的不整合の自動判別

グループに関する特徴量(Group)
・グループ化に使った検出ルール
・リンク数の分散
・グループ全体のリンク数との比

参照関係に関する特徴量(Relation)
・アンカー文字列に含まれる単語がリンク先文書のタイトル/見出し等に出現する割合

検出ルール	リンク数	リンク元 *s*	リンク先 *t*	アンカー文字列 *a*
link(s', t', a)	2000	URL_3	URL_1	製品α
link(s', t', a)	50	URL_4	URL_2	製品α

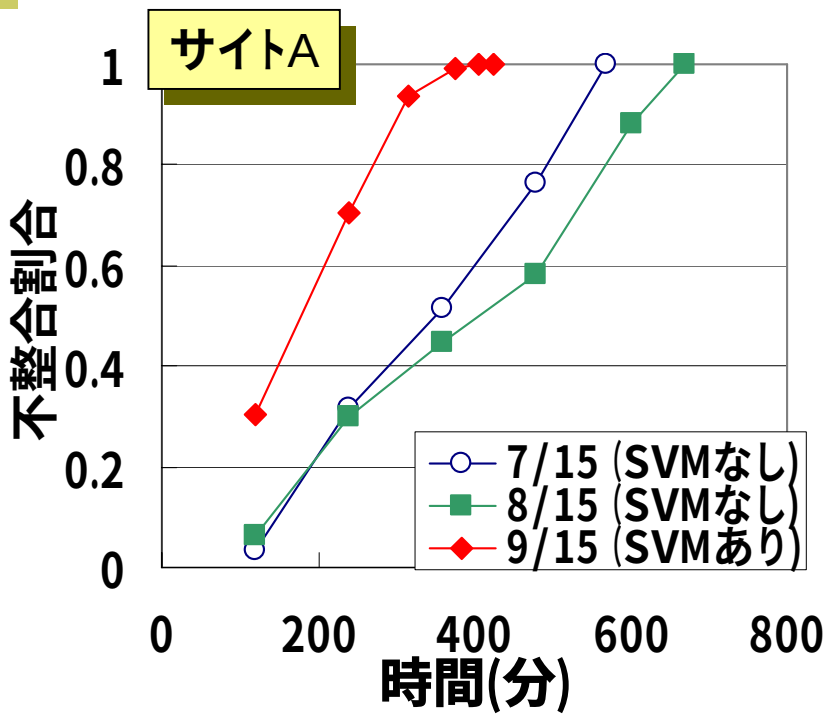
URLに関する特徴量(URL)
・URLに含まれるディレクトリ・ファイル名、およびその順番

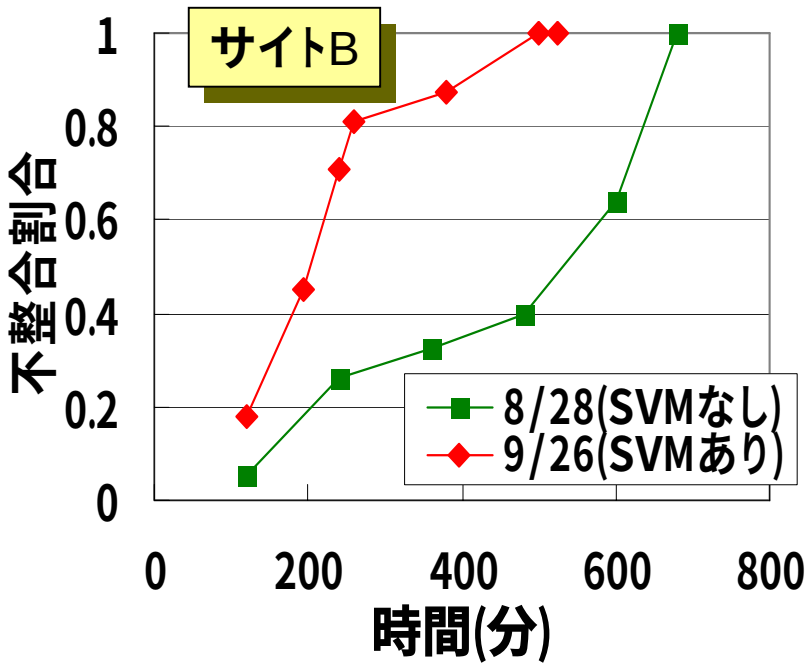
アンカー文字列に関する特徴量(Anchor)
・アンカー文字列に含まれる単語

3.3 論理的不整合の関連研究(2)

3.3.7 自動判別による検出効率向上

- SVMが「不整合」と判定したリンクを優先チェック
  - 初期段階で多くの不整合を検出→やがて飽和
  - 不整合総数の90％を見つけるまでに要する時間は約半分になった





発表の構成

- 1. 背景
- 2. Webサイト分析･診断手法
  - ユーザ指向の分析･診断手法
  - サイト指向の分析･診断手法
- 3. リンク整合性チェック
  - 物理的不整合
  - 論理的不整合
- 4. まとめ

4. まとめ

- Webサイト分析・診断手法について紹介
  - ユーザ指向の分析・診断手法
    - ユーザビリティテスト、アクセス/販売ログ解析、視線解析
  - サイト指向の分析・診断手法
    - サイト構造解析、ガイドライン検査、競合サイト比較、リンク整合性
- リンク整合性チェック
  - 血圧検査的なサイト診断手法
    - 簡単で、定期的にできて、潜在的な問題を発見可能
  - 物理的不整合
    - 古くから問題として指摘されており、自動修正の研究が盛ん
  - 論理的不整合
    - 不整合はたくさん存在するが、比較的未開拓のまま
    - 自動検出が第一の課題

参考文献(1)

[1] 2004 年度 全上場企業 ホームページ実態調査, 日興アイ･アール株式会社, 2004年11月, http://www.nikkoir.co.jp/rank/2004press.pdf

[2] 中小企業のIT（情報技術）活用状況等に関する調査, 商工中金, 2003年12月, http://www.shokochukin.go.jp/pdf/cb2003jyoho.pdf

[3] 日本の有力企業約250社のWebサイトの価値ランキング2004年版, 2004年7月, http://japanbrand.jp/we/1026/1.html

[4] Jules Yoshiyuki Tajima, プロフェッショナルWebプロデュース, SCC, 2001, 296p. (ISBN4-88647-222-2)

[5] COMPLEX, Webデザインワークフローガイド, エムディエヌコーポレーション, 2002, 191p. (ISBN4-8443-5623-2)

[6] ジェフリー･ヴィーン著, 長谷川憲絵訳, 戦うWebデザイン, エムディエヌコーポレーション, 2001, 242p. (ISBN4-8443-5596-1)

[7] (株)ビービット著, 篠原稔和監修, ウェブ･ユーザビリティルールブック, インプレス, 2001, 221p. (ISBN4-8443-1528-5)

[8] Steve Krug著, 中野恵美子訳, ウェブユーザビリティの法則, ソフトバンクパブリッシング, 2001, 215p. (ISBN4-7973-1597-0)

[9] ヤコブ･ニールセン著, 篠原稔和監修, グエル訳, ウェブ･ユーザビリティ, エムディエヌコーポレーション, 2000, 343p. (ISBN4-8443-5562-7)

[10] ヤコブ･ニールセン, マリー･タヒル共著, 風工舎訳, ホームページユーザビリティ, エムディエヌコーポレーション, 2002, 243p.(ISBN4-8443-5640-2)

参考文献(2)

[11] Jacob Nielsen, useit.com: Jakob Nielsen's Website, http://www.useit.com/

[12] Carol Barnum, Nigel Bevan, Gilbert Cockton, Jakob Nielsen, Jared Spool, Dennis Wixon, "The "magic number 5": is it enough for web testing?", in CHI '03 Extended Abstracts Conference on Human factors in Computing Systems, p.698-699, April 2003.

[13] Jason Hong and James A. Landay, "WebQuilt: A Framework for Capturing and Visualizing the Web Experience", in Proceedings of the 10th International World Wide Web Conference (WWW10), 2001.

[14] Jim Sterne ,Matt Cutler, "E-Metrics - Business Metrics For The New Economy", white paper, http://www.emetrics.org/articles/whitepaper.html

[15] Laura A. Granka, Helene A. Hembrooke, Geri Gay, Matthew K. Feusner, "Correlates of Visual Salience and Disconnect: An Eye-tracking Evaluation", Unpublished research report, Cornell University Human-Computer Interaction Lab, http://www.hci.cornell.edu/eyetracking/EyeTrackingsalience.pdf

[16] Ed Chi, Adam Rosien, Gesara Supattanasiri, Christiaan Royer, Amanda Williams, Celia Chow, "The Bloodhound Project: Automating Discovery of Web Usability Issues using the InfoScent Simulator", in Proceedings of ACM Conference on Human Factors in Computing Systems (CHI 2003), 2003.

[17] Bing Liu, Kaidi Zhao, Lan Yi, "Visualizing Web Site Comparisons", in Proceedings of the 11th International World Wide Web Conference (WWW2002), 2002.

[18] WatchFire WebXM, http://www.watchfire.com/products/webxm/

[19] Melody Y. Ivory, Marti A. Hearst, "Statistical Profiles of Highly Rated Web Sites", In ACM Conference on Human Factors in Computing Systems, CHI Letters, 2002.

[20] LinkScan, http://www.elsop.com/linkscan/

参考文献(3)

[21] Melody Y. Ivory, Marti A. Hearst, "The State of Art in Automating Usability Evaluation of User Interfaces", ACM Computing Surveys , Vol. 33 Issue 4, p. 470, 2001

[22] A. Chankhunthod, P. Danzig, C. Neerdaels, M. Schwarz, K. Worrel, "A Hierarchical Internet Object Cache", In Proceedings of USENIX Annual Technical Conference (USENIX'96), p.153, 1996.

[23] Graphics, Visualization, & Usability Center, GVU's Tenth WWW User Survey, Question 11, ”Problems Using the Web” ,1998. http://www.gvu.gatech.edu/user_surveys/survey-1998-10/graphs/use/q11.htm

[24] J. Pitkow. "Web Characterization Activity Answers to the W3C HTTP-NGs Protocol Design Group’s Questions". World Wide Web Consortium, 1998. http://www.w3.org/WCA/Reports/1998-01-PDG-answers.htm

[25] Seung-Taek Park, David Pennock, Lee Giles, Robert Krovetz, "Analysis of Lexical Signatures for Finding Lost or Related Documents" , Proceedings of the 25th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, p. 11, 2002.

[26] 中溝昌佳, 有山智洋, 森嶋厚行, 杉本重雄, 北川博之, "WWWにおけるリンク一貫性維持支援システムの開発", 電子情報通信学会第15回データ工学ワークショップ(DEWS2004), 2004.

[27] G. Buchanan, G. Marsden, H. Thimbleby, Meaningful Link Verification for Management and Maintenance of Web Sites, In Proceedings of the 8th International World Wide Web Conference (WWW8), Toronto, May 1999.

[28] 河合 英紀, 河野 泉, 石黒 義英, 福島 俊一, "サイト品質管理のためのリンク不整合検出", 電子情報通信学会第15回データ工学ワークショップ(DEWS2004), 2004.

[29] 河合 英紀, 河野 泉, 石黒 義英, 福島 俊一, "リンク不整合検出によるWebサイト診断 －論理的不整合の自動判定", 第66回情報処理学会全国大会, 3A-5, 2004.

[30] H. Ashman, "Electronic Document Addressing - Dealing With Change". ACM Computing Surveys, Vol. 32, Issue 3, p.201, 2000.