自然言語処理 129－8
（1999. 1. 21）

リンク情報を考慮したWeb検索システム

大森 貴博* 笹塚 清二* 水谷 正大†

* 東京情報大学 大学院 経営情報学研究科
† 東京情報大学 情報学科

抄 録

WWWページの検索効率の向上を目的とたキーワード群検索において，リンク情報を考慮した検索データベースを提案する．Webページで使われているタグの統計結果はリンク情報を考慮したシステムの必要性を示唆している．また，実際に構築した検索システムの評価実験を報告する．検索対象となるページ群へリンクしている後方ページ情報を考慮した検索システムは，検索対象ページ群からリンクしている前方ページ情報を考慮したシステムよりも，検索性能が高いことが分かった．

Web Search System considering Links among Web Pages

Takahiro Ohomori* Seiji Sasazuka* Masahiro Mizutani†

ohmori@rsch.tuis.ac.jp sasazuka@rsch.tuis.ac.jp, mizutani@rsch.tuis.ac.jp

* Graduate School, Tokyo Univ. of Information Sciences
† Dept. of Information Systems, Tokyo Univ. of Information Sciences

Abstract

A WWW search system considering the links among web pages is studied. Analysis of the Tag statistics in the web pages suggests the search database embedding the link information should be used to improve the seraching efficiency for the complex keywords search. Running this WWW search system, results shows that the efficiency of the system considering the backward-link pages from which the target pages are linked has a tendency to be better than one of the system considering the forward-link pages to which the target pages have links.

1 はじめに

Web検索システムの設計においては，クライアントからの真の要求の把握とその要求に見合った最適なWebページを見出す方法の開発が検索効率の向上のための理論的課題である．一方，システムの運用に際しては，収集したWebページ情報の肥大化にともない検索データベースの構築方法自体も大きな問題となる．つまり，頻繁に更新，登録されるWeb情報を反映するためにはデータベースも同じく頻繁に更新していかねばならない．

ここでは，検索データベースの更新が比較的容易な重み付き転置インデックスを拡張し，Webページ間のリンク情報を考慮した検索システムの構築を提案する．注目しているWebページに関するリンク情報には，そのページからリンクしている前方URLとそのページへリンクしている後方URL情報がある．本システムでは，この2つの情報をそれぞれ重みαとβで得点加算して検索データベースを構築する．

まず第2節では，リンク情報を考慮しない我々の研究を紹介する．検索データベースはWebページに登場する単語にタグ情報を考慮して得点を与えることによって構成されている．第3節では，リンク情報を考慮に入れるべき必要性をWebページで使われているタグの統計から論ずる．現在のタグの役割はページの論理構成のための利用から，実際にはページ整形のための利用へと大きく移行しており，単純にタグによる得点を考慮しただけの検索データベースでは検索効率の向上には限界がある．第4節では，URLで指定されるWebページの集合に関するいくつかの形式化を行う．第5節では，リンク情報を考慮して前方および後方ページからの重みαとβに依存した検索データベースを定義する．最後に，第6節で，検索システムにおける効率のパラメータ依存性に関する評価を報告し，考察を行なう．

2 従来の研究

クライアントからの検索要求を的確に捉え，これを受けて全文検索を行なうためには，本来，何らかの自然言語処理段階を経るシステムの構築が望ましい．しかし，Webページを処理対象とした場合には自然言語技術の不完全さ以前のさまざまな問題が指摘されている[1].

そこでWebページ検索システムを簡潔に構築する目的で，クライアントからの要求に見合うWebページを見出すための1つの方法としてWebページ内の単語ごとに得点を与えた転置インデックス法[2]を採用した．要求されたキーワード群を含むページをその得点和の順に並べ替え，高得点を持つページ順にクライアントに検索結果として返す図1のような検索システムを構築した．

2.1 Web検索システムの構成

クライアントからの問合せ文を受け取った検索カーネルはChaSen[3]にその形態素分析を要求し（処理a,b），その結果を受け取った検索カーネルは検索データベースに問合わせる（処理c,d）．最後に検索カーネルは検索データベースから受け取ったWebページのURL群をクライアントに回答として返すのである（処理e,f）．

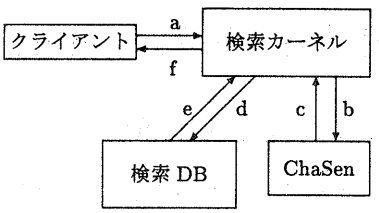

図1: Web検索システム

クライアントからの検索要求（処理a）として，複数の単語群{単語^1,...単語^N}のいくつかを含むURLページ集合に関するAND, OR, NOT演算，すなわち和（$\cup$），積（$\cap$），補（$\neg$）からなる集合演算を問合わせとして考えることにした．ただし検索対象が$\neg A$となるようなURL集合全体を参照するような場合は除外し，例えば，複合検索で$\neg A \cap B$が$B \cap \neg A = B \setminus A$と書けるように，単項演算子$\neg$を含む式が$B$から集合$A$の要素を取り去るという集合差として解釈できる場合を考えることにする．また，本システムは検索要求としてテキストを与えたとしても，それがChaSenにわたされ形態素分析されて複合ANDキーワード検索として働くようにしている．

2.2 検索データベースの構成

Web ロボットが収集した Web ページ群は，まず次の項目をフィールドとしたレコード形式にしたっがって蓄積され，検索データベース構築のための 1 次情報として利用される．ただし，最後の項目" リンク先の URL 集合" はリンク情報を考慮した検索データベース構築のために追加されるフィールドである．

- ◇ Web ページの URL
- ◇ <title>で指定されたページタイトル
- ◇ 表示データ（<meta contents> + 本文先頭部）
- ◇ Web ページの全文データ
- ◇ Web ページのサイズ
- ◇ Web ページから他ページに張られているリンク数
- ◇ （リンク先の URL 集合）

検索データベースの構築には次の段階を経て行われた．Web ページ内の得点計算の対象となる単語の切り出しには形態素分析器 ChaSen を使い，「名詞、動詞、形容詞」に分類されたものを検索用の単語として登録する．また，各 Web ページ内の単語得点の計算にはページ内の単語登場頻度に加えて，表 1 のように単語を囲む HTML タグにもその意味合いに応じた重み $\mathcal{S}$ を与えて得点に加算する方式を用いた．HTML タグに重みをつけるやり方は全文検索システム Namazu[4] でも採用しており，論理タグだけに重みをつける方式をとっている．

こうして，Web ページごとに単語得点を計算しておき，収集した Web ページ群から次のようなレコード構造

- ▷ 単語
- ▷ (URL ページ$_1$, 得点$_1$)
- ⋮
- ▷ (URL ページ$_m$, 得点$_m$)

を持った単語を索引キーとする**得点付けされた検索データベース** $\mathcal{DB}_\mathcal{S}$ を作成する．この方法を TF.IDF 法 [2] による得点づけと比べると，ページ単体のみで得点評価できるために計算量は格段に少なくなる．また，タグによる重みを考慮せずに（$\mathcal{S}=0$）作成した検索データベースを $\mathcal{DB}_0$ と記す．

検索要求から得られたキーワード群に対して，この $\mathcal{DB}_\mathcal{S}$ を探索し，高得点順に URL ページ群を結果として返すことによって Web 検索システムを構築した．

表 1: タグに囲まれた単語の追加得点 $\mathcal{S}$

タ　グ	得点
<META NAME="robot" KEYWORDS=...>	10
<TITLE>	5
<H1>	7
<H2>	6
<H3>	5
<H4>	4
<H5>	3
<H6>	2
<FONT SIZE=7>	5
<FONT SIZE=6>	4
<FONT SIZE=5>	3
<FONT SIZE=4>	2
<B>	2
<STRONG>	2
<I>	2
<U>	2
<A HREF=>…</A>	2
<IMG SRC= ALT="…">	2

3 リンク情報の必要性

本研究の目的は，このようにして構築された Web 検索システムを改良し検索効率の向上を図るための方法の 1 つとしてリンク情報を考慮したシステムを検討することにある．

従来の研究では Web ページにおけるタグの役割を重視し，その役割に応じた重みを単語の得点計算の際に加算していた．HTML ver.2.x を経て HTML ver.3.x の普及，さらに HTML ver.4.0 への移行にともない，タグの使われ方に大きな変化が現れている．この節で述べるように，Web ページ構造の論理的指定のためのタグ利用から，ページ表示レイアウト制御のためのタグ利用へとタグの使われ方の変化傾向が伺える．この傾向を加速させているのが Web ページ作成ツールの高性能化とその普及である．こうなると，たとえば "<H1>..</H1>のような見出しタグに使われる単語は Web ページ内で一定の重みを持つべきである" というような素朴な前提自体を見直さざるを得ない．実際，そのようなツールで作成した Web ページには

<META NAME="generator" CONTENT="creater 名">

のように META タグ内に generator が現れる．

表 2 は jp 第 2 レベル・ドメインそれぞれにおける代表的なタグの利用頻度を集計したものである．こ

表 2: ドメイン別の規格化されたタグ平均出現数/ページ

	ac.jp	ne.jp	co.jp	or.jp	go.jp	gr.jp	ALL
<TITLE>	1.24	0.73	0.92	0.93	0.79	1.39	0.95
<H1>	0.54	0.14	0.16	0.21	0.19	0.13	0.22
<H2>	0.70	0.27	0.35	0.36	0.42	0.29	0.39
<H3>	0.80	0.45	0.28	0.49	0.54	0.28	0.47
<H4>	0.44	0.13	0.11	0.19	0.24	0.19	0.20
<H5>	0.06	0.05	0.03	0.04	0.04	0.01	0.04
<H6>	0.01	0.02	0.00	0.03	0.01	0.01	0.01
<FONT SIZE=4>	0.73	0.73	0.64	0.98	0.97	0.40	0.77
<FONT SIZE=5>	0.35	0.33	0.38	0.44	0.25	0.17	0.32
<FONT SIZE=6>	0.13	0.11	0.07	0.13	0.10	0.04	0.10
<FONT SIZE=7>	0.04	0.03	0.02	0.04	0.01	0.03	0.03
<I>	0.52	0.44	0.57	0.50	0.47	1.04	0.56
<EM>	0.11	0.14	0.18	0.10	0.05	0.11	0.12
<B>	3.48	4.69	5.43	4.36	3.42	3.64	4.23
<STRONG>	0.40	0.83	0.56	0.46	0.20	0.27	0.48
<U>	0.10	0.07	0.09	0.14	0.20	0.05	0.11
<A HREF>	13.7	11.5	9.36	13.5	7.28	6.99	10.4
<IMG SRC=>	7.45	6.36	8.09	7.86	4.18	3.24	6.26
<TABLE>	1.87	2.47	3.03	2.59	1.75	1.17	2.22
<TR>	7.24	9.16	8.76	8.69	6.65	4.82	7.77
<TD>	20.8	19.8	19.7	20.8	23.1	11.4	19.8
<ADDERSS>	0.23	0.04	0.05	0.1	0.07	0.03	0.08
<PRE>	0.47	1.44	1.08	0.41	0.20	0.88	0.77

のために，ドメイン d 毎に同じ数の Web ページを収集して，各ページのファイルサイズの合計 S_d，利用されているタグ T 毎の総数 n_d^{T}，および全ドメインでのページ当たりの平均ファイルサイズ $\langle S \rangle$ を求めた．それらから平均ファイルサイズ当たりの換算ページ数 $\bar{P}_d = S_d/\langle S \rangle$ をドメイン毎に算出して，規格化されたタグ T の平均出現数 $n_d^{\mathrm{T}}/\bar{P}_d$ を計算したものが表 2 である．

表 2 から次のようなタグの利用傾向が見てとれる．

- ac.jp ドメインでは他のドメインに比べて<H1>から<H4>の見出しタグの利用率が高い．
- <EM>，<STRONG>や見出しタグよりも<I>，<B>や<FONT SIZE>などのように，論理的タグよりも直接的にページ整形するタグの利用率が高い
- <TABLE>タグが多用され<TR>や<TD>の利用率が高く，表はネストされて使われているかもしれない．

こうした傾向は，手入力で文書の論理構造に応じたタグを書くといった‘古典的 HTML スタイル’から，何らかのツールを使った Web ページ生成の増加を示唆していると考えられる．実際，全ドメインを通じて全ページ数の 15%以上に generator を含む<META ...>があり，Web ページ作成ツールの利用が進んでいることが分かった．

したがって，Web ページ検索において，ac.jp ドメインでは表 1 で見られるようなタグ（特に<H1~3>）による単語の重み付けを考慮することに意味があるが，総じて文中でのタグの論理的意味合いが減少しており，タグによる重み付けをことさら重視するだけでは検索効率の向上に限界があると推測できる．

そこで，本研究では第 5 節で述べる方法を使って，Web ページ内の単語得点の評価にリンク先の前方ページおよびリンク元の後方ページにある当該単語からの重みを考慮することで検索効率の向上を考えることにした．

4 URL ページ集合

URL で指定される Web ページを URL ページと呼ぶ．URL ページ u 内でタグ<A HREF=..>または<FRAME SRC=..>が使われているとき前方 URL ペー

ジの集合を考えることができ

$$\mathcal{L}u = \{u \text{からリンクされている URL ページ}\}$$

によってリンク先作用素 $\mathcal{L}$ を定義する．また，URL ページ u に着目したとき u にリンクしている後方 URL ページの集合を考えることができ

$$\mathcal{L}^{-1}u = \{v \mid u \in \mathcal{L}v\}$$

によってリンク元演算子 $\mathcal{L}^{-1}$ を定義する．URL ページ集合 U が $U = \mathcal{L}U = \{\mathcal{L}u \mid u \in U\}$ であるとき U を $\mathcal{L}$ 不変，また $U = \mathcal{L}^{-1}U$ のとき U を $\mathcal{L}^{-1}$ 不変と呼ぶ．$\mathcal{L}$ 不変かつ $\mathcal{L}^{-1}$ 不変な URL 集合を極大と呼ぼう．

いま，URL ページ集合 U に関して，前方 $\mathcal{L}$ 境界 ∂_+U を

$$\partial_+U = \{u \in U \mid \mathcal{L}u \setminus U \neq \phi\},$$

後方 $\mathcal{L}$ 境界 ∂_-U を

$$\partial_-U = \{u \in U \mid \mathcal{L}^{-1}u \setminus U \neq \phi\}$$

で定義する．すなわち，∂_+U は U の部分 URL ページ集合で，そこからリンクされている前方 URL ページ群が集合 U に含まれていない URL ページ集合，∂_-U は U の部分 URL ページ集合で，そこへリンクしている後方 URL ページ群が集合 U に含まれていない URL ページ集合である．したがって，

$$\mathcal{L}(U \setminus \partial_+U) \subseteq U, \quad \mathcal{L}^{-1}(U \setminus \partial_-U) \subseteq U$$

であることに注意する．

とくに混乱のない限り URL とその URL ページを同一視するが，URL 自体と URL を指定して入手した URL ページとを区別して考えねばならないときもある [5]．たとえば，URL ページ集合 U に対し，$|\mathcal{L}U|$ でその前方 URL ページの総数を，$\|\mathcal{L}U\|$ で U から $\mathcal{L}U$ への前方リンクしている前方 URL の総数を表わす．$\mathcal{L}U$ にある URL ページには U 内から 1 つ以上の URL によって前方リンクされているため，一般に $|\mathcal{L}U| \leq \|\mathcal{L}U\|$ であることに注意する．同様に，$\|\mathcal{L}^{-1}U\|$ を後方リンク総数，つまり U に向かって前方リンクする $\mathcal{L}^{-1}U$ からの URL の総数とすると，$|\mathcal{L}^{-1}U| \leq \|\mathcal{L}^{-1}U\|$ である．

本研究で考えるようなリンク情報を考慮した検索システムを構築するためには，理論的には検索対象とする URL ページの集合 U は極大でなければならない．極大な URL ページ集合を見出すことは，世界中の URL ページ集合以外には，一般にはきわめて困難である．理論的には検索評価の対象外となる境界ページ集合 $\partial_\pm U$ から U の外部へリンクしている URL ページ集合 $\mathcal{L}U \setminus U$ および U の外部から U へリンクしてくる URL ページ集合 $\mathcal{L}^{-1}U \setminus U$ は

$$\lim_{|U|\to\infty} \frac{|\mathcal{L}U \setminus U|}{|U|} \to 0,$$

$$\lim_{|U|\to\infty} \frac{|\mathcal{L}^{-1}U \setminus U|}{|U|} \to 0$$

となる．したがって，十分大きな U をとる限り URL 集合 $(\mathcal{L}U \cup \mathcal{L}^{-1}U) \setminus U$ は U に比べて相対的に小さくなり，実質的にシステムの運用には影響がでないと期待することができる．ただし，具体的に URL 集合 U を与えて，この推測を検証することは容易ではない．

本研究では，選び出した URL 集合 U に関して厳格に正しいリンク情報を与えない U 内の境界 URL 集合であっても，$\partial_\pm U \cap U \neq \phi$ を満たすような URL ページであれば，リンク情報を考慮する検索データベースに登録することにした．

5 検索データベースの構築

本研究では，リンク情報を考慮した検索システムの構築のために，第 2.2 節で得点付けられた検索データベース $\mathcal{DB}_S$ における索引単語の各得点を，次のように再定義することによって得られたものを**リンク情報を考慮した検索データベース** $\mathcal{DB}_S(\alpha,\beta)$ とする．

ある URL ページ u とそこからリンクされている前方 URL ページ群 $\mathcal{L}u = \{f_1, f_2, \ldots f_m\}$，および u へリンクしている後方 URL ページ群 $\mathcal{L}^{-1}u = \{b_1, b_2, \ldots b_n\}$ を考える．第 2 節で述べた方法で，ページ u での単語群 $W_u = \{w^1, w^2, \ldots, w^N\}$ に対応する得点 $S_u = \{s(w^1), s(w^2), \ldots, s(w^N)\}$ は既に計算されている．同様に，ページ u からの前方 URL ページ f_i での単語群 $W_{f_i} = \{w^1_{f_i}, w^2_{f_i}, \ldots, w^I_{f_i}\}$ の得点 $S_{f_i} = \{s(w^1_{f_i}), s(w^2_{f_i}), \ldots, s(w^I_{f_i})\}$，およびページ u への後方 URL ページ b_j での単語群 $W_{b_j} = \{w^1_{b_j}, w^2_{b_j}, \ldots, w^J_{b_j}\}$ の得点 $S_{b_j} = \{s(w^1_{b_j}), s(w^2_{b_j}), \ldots, s(w^J_{b_j})\}$ も既に得られているとしよう．このとき，リンクを考慮した

URL ページ u 内の単語 w^k の得点 $s^{(\alpha,\beta)}(w^k)$ を改めて次のように定義する．

$$s^{(\alpha,\beta)}(w^k) = s(w^k) + \alpha \sum_{i=1}^{m} s(w_{f_i}^k) + \beta \sum_{j=1}^{n} s(w_{b_j}^k)$$

但し，$w_{f_i}^k = \{w^k\} \cap W_{f_i}$ はページ u の単語 w^k であって且つ u からリンクしている前方ページ $f_i (i = 1, \ldots m)$ に含まれる単語，同様に，$w_{b_j}^k = \{w^k\} \cap W_{b_j}$ はページ u の単語 w^k であって且つ u へリンクしている後方ページ $b_j (j = 1, \ldots n)$ に含まれる単語である（そのような単語がないときは重みを加算しない）．URL ページ u 内の単語の得点として，ページ u 内での得点に加えて，リンク先の前方ページにある単語の得点を重み α で，リンク元である後方ページにある単語の得点を重み β で加算していることに注意する．

こうして得られたリンク情報を考慮した得点付けされたデータベース $\mathcal{DB}_{\mathcal{S}}(\alpha,\beta)$ は $\mathcal{S}$ を留めるごとに $\mathcal{DB}_{\mathcal{S}}(0,0) = \mathcal{DB}_{\mathcal{S}}$ を満たす α と β に関する 2-パラメータ族をなしている．このとき問題となるのは，検索効率のパラメータ依存性である．

6 性能評価と考察

リンク情報を考慮した検索システムとして第5節で定義したデータベース $\mathcal{DB}_{\mathcal{S}}(\alpha,\beta)$ を使ったシステムを構築して，検索性能に関する評価実験を行なった．

表 3: URL ページ数

CPU∧ 価格 ∧ 相場 ∧Celeron		
$\|V\|$	$\|\mathcal{L}V\|$	$\|\mathcal{L}^{-1}V\|$
107	7,925	438

ピザ ∧ ワイン ∧ レストラン ∧ 千葉		
$\|V\|$	$\|\mathcal{L}V\|$	$\|\mathcal{L}^{-1}V\|$
96	14,692	1,312

実際の性能評価は次のように行われた．まず，適当な検索キーワード群によってあらかじめ選び出した URL ページ集合を V，V の各ページからリンクされている前方 URL 集合を $\mathcal{L}V$，V へリンクしている後方 URL ページ集合を $\mathcal{L}^{-1}V$ とする．第4節で述べたように，$\mathcal{L}^{-1}V$ を正確に求めることは一般に困難であるが，ここでは V を含む十分に大きな URL 集合 $\mathcal{U}$ から V の各ページへリンクしている URL ページ群全体を近似的に $\mathcal{L}^{-1}V$ と見なした．

表 3 は，今回の評価のために AND キーワードによって絞り込んで選び出して実験評価した中の例で，検索対象となる URL 集合 V について検索対象ページ数 $|V|$ とその前方 URL 数 $\|\mathcal{L}V\|$ および後方 URL 数 $\|\mathcal{L}^{-1}V\|$ を示している．表 3 のように，検索対象となる URL ページ集合を決めると，その前方リンク数に比べて後方リンク数はかなり少ないという傾向が得られた．

収集してくる U ページ集合に関して，$|U|$ の大きさが比較的小さい間はその前方 URL 集合数 $\|\mathcal{L}U\|$（リンク数）は URL ページ数 $|U|$ よりも速く増加し，U がある程度大きくなって $\mathcal{L}$ 不変集合に匹敵するようになると前方 URL 集合の増加率は低くなり，$\mathcal{L}$ 不変になった段階で両者の数は一致する．一方，後方 URL 集合 $\mathcal{L}^{-1}U$ を見出していくには十分大きく U を収集する必要があり，U の増加に比べて後方リンク数 $\|\mathcal{L}^{-1}U\|$ の増加率は低くいと考えられる [6]．表 3 が示すような一般的傾向は，この考察を傍証している．

次に，検索システムの性能のパラメータ依存性を調べるために，データベース $\mathcal{DB}_{\mathcal{S}}(\alpha,\beta)$ に関して，第 2.2 節で論じたようにタグの重みを考慮しない場合と考慮した場合のそれぞれに対し，パラメータ (α,β) の組 $(0,0)$, $(1,0)$, $(0,1)$, $(1,1)$ について合計 8 通りの方法を評価した．

表 3 のキーワード群に関する結果が表 4 である．表には，データベース $\mathcal{DB}_{\mathcal{S}}(\alpha,\beta)$ を使って検索対象となる URL 集合 V を得点順に並べ，上位 1 ～ 10 位，11 ～ 20 位，21 ～ 30 位，31 ～ 40 位のそれぞれの中から被験者が検索対象としてふさわしいと判断した個数の平均値，および 8 通りの検索方法の中で良いと印象をもった上位 3 つの方法の平均得票率を示した．

表 3 およびその他の調査結果から，タグによる重み付けをしない（$\mathcal{S} = 0$）検索データベースを利用した場合では

- リンク元を考慮した方法（$(\alpha,\beta) = (0,1)$）は検索効率が高い．
- リンク先を考慮した方法（$(\alpha,\beta) = (1,0)$）は

表 4: 評価結果

CPU∧価格∧相場∧Celeron							
$\mathcal{DB}_{\mathcal{S}}(\alpha,\beta)$	1 ~ 10	11 ~ 20	21 ~ 30	31 ~ 40 (位)	第 1 評価	第 2 評価	第 3 評価 (%)
$\mathcal{DB}_0(0,0)$	5.6	5.6	3.8	5.8	0	17	8
$\mathcal{DB}_0(1,0)$	4.3	4.8	3.3	4.4	0	0	0
$\mathcal{DB}_0(0,1)$	6.7	9.9	7.8	7.5	33	50	0
$\mathcal{DB}_0(1,1)$	3.3	6.8	8.0	6.5	0	0	25
$\mathcal{DB}_{\mathcal{S}}(0,0)$	7.0	6.5	8.1	7.0	0	8	25
$\mathcal{DB}_{\mathcal{S}}(1,0)$	3.9	3.8	4.5	8.5	0	0	0
$\mathcal{DB}_{\mathcal{S}}(0,1)$	6.8	10	9.8	6.6	50	17	17
$\mathcal{DB}_{\mathcal{S}}(1,1)$	4.1	5.3	9.9	10	8	0	17

ピザ∧ワイン∧レストラン∧千葉							
$\mathcal{DB}_{\mathcal{S}}(\alpha,\beta)$	1 ~ 10	11 ~ 20	21 ~ 30	31 ~ 40 (位)	第 1 評価	第 2 評価	第 3 評価 (%)
$\mathcal{DB}_0(0,0)$	1.0	0.50	0.33	0.17	0	0	0
$\mathcal{DB}_0(1,0)$	1.0	0.75	0.67	0.08	8	8	8
$\mathcal{DB}_0(0,1)$	1.4	0.50	0.33	0.17	8	17	17
$\mathcal{DB}_0(1,1)$	1.2	1.3	0.17	0	8	8	0
$\mathcal{DB}_{\mathcal{S}}(0,0)$	1.9	0.58	0.33	0	25	17	8
$\mathcal{DB}_{\mathcal{S}}(1,0)$	1.9	0.83	0.08	0	8	17	8
$\mathcal{DB}_{\mathcal{S}}(0,1)$	1.9	0.92	0.08	0	0	0	25
$\mathcal{DB}_{\mathcal{S}}(1,1)$	1.9	0.75	0.08	0	8	0	0

効率的でない.

- リンクを考慮しない方法 ($(\alpha,\beta)=(0,0)$) はリンクを考慮した方法よりも良い結果を示している.
- リンク先とリンク元の両方を考慮した方法 ($(\alpha,\beta)=(1,1)$) には著しい効果はない

という傾向があることが分かった. また, タグによる重み付けをした検索データベースを利用した場合 ($\mathcal{S}\neq 0$) では

- 全般的に重み付けをしない場合と同様な傾向が見られる
- 検索結果の適合性は重みを付けた方が勝っている

という傾向にあった.

したがって, 本研究の調査範囲では, リンク先情報を考慮した検索システムの性能は高いものではなく, リンク元を考慮した検索システムが有利であることが分かった.

リンク先情報を考慮すべき理由として, そのWebページを書いている本人がリンク先の内容を知っており, それゆえリンク先のページは現在のページ内容に直接の関係があるはずだという論拠を想定することができる. しかしながら, リンク先情報は必ずしも検索要求を反映しない場合があることを評価結果が示している.

このリンク先情報を考慮したシステムの検索効率の低下傾向は次のように考えることができる. 第5節で述べた方法では, 検索対象ページが含んでいる複合キーワードのどれか1つのキーワードを含むページからであっても得点加算が可能であるために, 検索主題から逸脱したページの得点が高くなり得て, 結果的に本来望んでいる検索ページの順位が低下することになったと理解できる. 逆にいえば, 関連情報の取得のためにはリンク先情報を考慮した検索システムが適切となる. 実際, リンク先情報を考慮した方法ではリンク集を掲載したページが上位に集まる傾向が見られた.

一方, リンク元を考慮した検索システムが高い評価傾向を示した理由として, リンクを張るときには本人がそのページ主題を知っているために, 後方ページ群からの得点寄与は検索対象となるURLページ群に含まれている複合検索キーワード群を多く含むときに大きくなり, その結果として検索順位が上がると考えることができる. 検索主題がはっきりしている場合には, リンク元情報を考慮した検索システム

が適切となる.

参考文献

[1] 渡辺日出雄,「Web 文書に対する言語処理の問題点と言語処理を援助するタグセットについて」, 情報処理学会研究報告 98-NL-127,pp95-100.

[2] 長尾真, 黒橋禎夫, 佐藤理史, 池原悟, 中野洋, 『言語情報処理』(言語の科学 第 9 巻), 岩波書店 (1998 年).

[3] 形態素解析器 ChaSen (茶筌), 奈良先端化学技術大学自然言語処理学講座.

[4] 高林 哲,「全文検索システム Namazu」, http://saturn.aichi-u.ac.jp/~ccsatoru/Namazu/intro.html.en

[5] 来住伸子,「分野を特定した自動収集による WWW 情報検索」, 情報処理学会研究報告 98-NL-124,pp87-94.

[6] 来住伸子, Private Communication