Evaluation of Web Retrieval Methods Using Anchor Text

Kenji TATEISHI, Hideki KAWAI, Susumu AKAMINE, Katsushi MATSUDA
and Toshikazu FUKUSHIMA
Internet Systems Research Laboratories, NEC Corporation
8916-47, Takayama-Cho, Ikoma, Nara 630-0101, Japan
{tateishi,kawai,akamine,mat,fuku}@hml.cl.nec.co.jp

Abstract

In this paper, we evaluate two types of anchor texts: a page anchor and a site anchor. Since the anchor text tends to summarize information referred ahead, it can be expected that the terms appearing there have important meaning in information retrieval. We introduce a retrieval method to give high priority to the terms in the anchor text. In the experiment, we compared the proposed method with the base line which indexed only page documents. The result indicated that both methods had almost the same accuracy, and that there were many queries in which the accuracy much differed between two methods. It can be expected that the improvement on the queries in which the proposed method was inferior to base line will be achieved by deleting overlapped anchor texts toward the same page.

1. Introduction

The text in a link is called anchor text. Since the anchor text tends to summarize information referred ahead, it can be expected that the terms appearing there have an important role in information retrieval. We participated in this NTCIR-WEB task to clarify the effect of indexing them. We introduced not only conventionally used page anchors, but also site anchors, and used the retrieval method to give high priority to the terms in anchor texts. Hereafter, we discuss the two types of anchor texts, the retrieval method and the evaluation result.

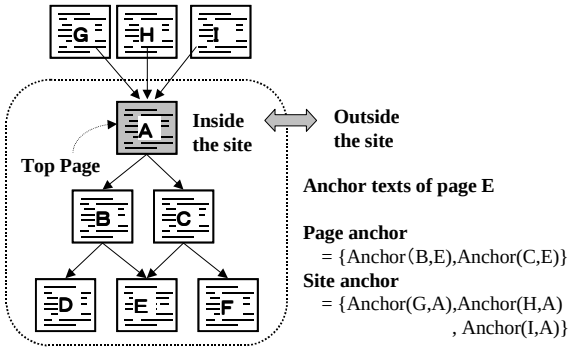

Figure 1. An example of the page anchor and the site anchor

2. Anchor text

We used the following two types of anchor texts for information retrieval.

(1) An anchor text which summarizes content of a web page (hereafter, page anchor)
(2) An anchor text which summarizes content of a web site (hereafter, site anchor)

First of all, the page anchor is the text in a link to a given web page. For instance, in Figure 1, page anchors of page “E” are equal to {Anchor (B, E), Anchor(C, E)}. Here, Anchor(x,y) shows an anchor text in a link of page x to page y. This page anchor has the same definition as a so-called usual anchor text.

Next, the site anchor is the text in a link to the top page of a given web site. For instance, {Anchor(G,A), Anchor(H,A),Anchor(I,A)} correspond to the site anchors of page “E” in Figure 1. Since the web site is usually constructed under the assumption of visitors browsing each page via the top page, the expressions shared on the entire web site tend to be omitted. For instance, suppose a web site where gourmet

© 2003 National Institute of Informatics

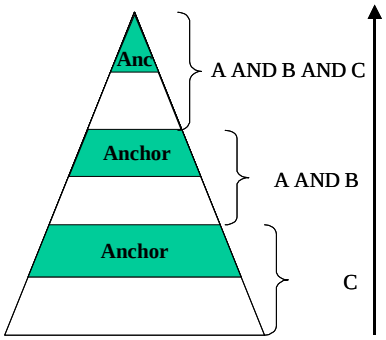

Figure 2. An image of the retrieval model (For three query terms)

information on an entire country is treated. There, "Gourmet" and "Restaurant retrieval" are often shown as anchor texts of the main page while minimal information is attached to the anchor texts of the internal pages such as "Kansai", "Chinese cooking" and "Nara". Therefore, not only page anchors but also the site anchor have to be taken in order for the anchor text to summarize the entire content of the web page.

3. Retrieval method

We explain the retrieval method used to evaluate two types of anchor texts in Section 2.

3.1 Decision of the top page for the site anchor

The top page of the web site where a certain page E belongs means a page that represents the whole web site or a certain area of it. It must fulfil the following three conditions.

(1) The page has the same domain as E.
(2) The page is hierarchically above E.
(3) The number of external (from different domain) links is more than *n*.

According to these requirements, we assume that a top page has many external links. From here on we set it as 2.

3.2 Ranking method

We use a ranking method that gives high priority to the anchor text, which processes according to the following.

Step 1: Decrease web pages by using the AND function made from all query terms. Each query term can exist either in the web page document or in the anchor texts of the page.

Step 2: To the web pages narrowed in Step 1, give high priority to web pages whose anchor texts contain the terms.

Step 3: Delete one query term, and repeat from Step 1.

Thus, it is intuitively the pyramidal ranking shown in Figure 2, and the web pages whose anchor texts contain the same kind of query terms can attain a high position in the ranking.

In step 1, we reduced web pages beforehand with the AND function because the preliminary experiment informed us it was better than the simple model using only the step 2 process.

In step 2, we calculate the sum of tf * idf over all query terms for the anchor texts only, and simply multiply tf by idf without any normalization of that length. The tf means the frequency of a certain term in the anchor texts. When a web page has multiple anchor texts, we connected them into one and counts tf.

In step 3, when TopicPart is a Title, the rightmost query term is deleted because it was provided beforehand that the query term on the right hand side is more important for the NTCIR-Web task. On the other hand, when TopicPart is a Description, we delete a query term whose idf is minimal.

4. Evaluation

4.1 Comparison systems

We compared the following three kinds of retrieval system.

(1) The base line system that indexes only web page documents (hereafter, Baseline)

Table 1. Experimental results (upper table: TopicPart = Title, below table: TopicPart = Description)

Title	10G				100G			
	Prec@10	Prec@20	R-Pre	Ave Pre	Prec@10	Prec@20	R-Pre	Ave Pre
Baseline	0.2298	0.2021	0.1862	0.1398	0.3213	0.3106	0.2007	0.1246
Page Anchor	0.2413	0.2043	0.1815	0.1420	0.3234	0.3032	0.1922	0.1297
Page+Site Anchor	0.2391	0.2043	0.1815	0.1408	0.3149	0.2862	0.1859	0.1285

Description	10G				100G			
	Prec@10	Prec@20	R-Pre	Ave Pre	Prec@10	Prec@20	R-Pre	Ave Pre
Baseline	0.2660	0.2043	0.1835	0.1469	0.3149	0.2947	0.2040	0.1229
Page Anchor	0.2511	0.1856	0.1561	0.1371	0.2652	0.2554	0.1716	0.1189
Page+Site Anchor	0.2511	0.1856	0.1561	0.1371	0.2587	0.2478	0.1713	0.1178

(2) The retrieval system in Section 3 that uses the page anchor as the anchor text (hereafter, Page Anchor).

(3) The retrieval system in Section 3 that uses the page anchor and the site anchor as the anchor text (hereafter, Site Anchor)

We implemented a base line system which used a ranking method of Okapi[2]. Okapi is a retrieval method based on the probabilistic model, and a system using this method has given successful results in past TRECs. Note that since Baseline was not submitted to the evaluation, it was not included in the pooling web pages. The experiment was conducted with both a 10G and a 100G index The evaluation scales are Prec@10, Prec@20, R-Precision and Average Precision.

From all these systems, we selected the character basis indexing style and used "Chasen"[1] for morphological analysis when TopicPart was the Description. The morphemes we took as the query terms were noun, verbal noun, and adjectival noun. Among these morphemes, those mutually adjacent were connected and treated as one term. We confirmed by a preliminary experiment that higher accuracy could be obtained when the mutually adjacent morphemes were connected rather than not. Moreover, we excluded some noun terms as stop words. For example, 情報(information), 説明(explanation), 文書(document), and 関心(concern), etc.

4.2 Experimental result

The experiment results are shown in Table 1 (upper table: Title, below table: Description). When TopicPart was the Title, both Page Anchor and Baseline have almost the same performance, although Page Anchor slightly exceeded Baseline in some of the evaluation scales. The accuracy of Site Anchor was inferior to that of Page Anchor in all the evaluation scales. On the other hand, when TopicPart was the Description, neither Page Anchor nor Site Anchor performed as well as Baseline. Therefore, from this result, we could not observe improvement in the method, which gave high priority to the terms, which appear in the anchor texts. Past TRECs have also reached similar conclusions[3], so this result supports them.

4.3 Discussion

Table 2 shows the comparison of Baseline and Page Anchor results for each query when TopicPart is the Title and the evaluation scale is Prec@10. There were many queries in which the accuracy much differed between two systems. There were 17 of 47 queries queries in which Page Anchor was inferior to Baseline.

When we put attention to this case, there were two reasons. The one was because the query terms , first of all, hardly hit in the anchor texts and Page Anchor could not make use of its feature. 4 of 17 quries were

{各種資格試験,各種資格試験,各種資格試験 各種資格試験シリーズ,各種資格試験,各種資格試験,各種資格試験,各種資格試験,各種資格試験,各種資格試験,各種資格試験}

Figure 3. An example of overlapped page anchors "資格試験 (Qualifying exam)" when the query is "資格試験,情報処理,IT(Qualifying exam, information processing, IT)"

adopted to this case. This result can be found from the line of "Anchor Hit" in Table 2. "Anchor Hit" means the number of web pages within the top 10 which more than one query terms apper in the anchor texts. Thus, for example, "サルサ,学ぶ,方法 (Salsa, Learn, methods)" and "速読法,効果 (Speed reading method, effects)" had few web pages that hit query terms in the anchor texts.

Next, the another case was because there were many overlapped anchor texts towad the same web page and then Page Anchor gave higher score to the web page than should have given. These strongly correspond to the ones used for navigation in the same web site. Figure 3 illustrates an example of overlapped page anchors "資格試験 (Qualifying exam)" when the query is "資格試験,情報処理,IT (Qualifying exam, information processing, IT)". This problem can be solved by deleting the overlapping anchor texts toward the same web page from the same web site.

However, we don't know exactly whether the aforementioned problem represents all problems of accuracy decline, though it is also an important problem The proposed system and Baseline differs in that it decreases web pages with the AND function in addition to using anchor text. The proposed system didn't performed as well as Baseline when TopicPart was the Description (See Table 2) was that morphological analysis caused the system to retain many unnecessary query terms, then the AND function decreased web pages too much. Therefore, we plan to evaluate the proposed system again by preparing another base line system which is equal to the proposed system at the point of the AND function.

5. Conclusion

In this paper, we evaluated two types of anchor texts: page anchor and site anchor, for verifying the effect of indexing the anchor text. We also introduced a retrieval method to give high priority to the terms in the anchor text. . In the experiment, we compared the proposed method with the base line which indexed only page documents. The result indicated that both methods had almost the same accuracy, and that there were many queries in which the accuracy much differed between two methods. It can be expected that the improvement on the queries in which the proposed method was inferior to base line will be achieved by deleting overlapped anchor texts toward the same page. In future works, we plan to solve this problem and to make a fairer evaluation compared with the baseline system. In addition, we will develop a new type of the anchor text which includes site path anchor texts from the top page to a certain page (for example, Anchor(A,C) and Anchor(C,E) in Figure 1) in addition to page anchors and site anchors.

References

[1] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, O. Imaichi, T. Imamura, *Japanese morphological analysis system "Chasen"*, Available from http://cactus.aist-nara.ac.jp/lab/nlt/chasen/distribution.html

[2] S.E.Robertson, S.Walker. *Okapi/Keenbow at TREC-8*, The Eighth Text Retrieval Conference, 1999.

[3] David Hawking, *Overview of the TREC-9 Web Track*, The Ninth Text Retrieval Conference, 2000.

Table 2. The comparison of Page Anchor and Baseline at Prec@10 by each query

Query ID	Anchor Hit	Keyword	Baselin	Page Anchor	Diff
8	0	サルサ,学ぶ,方法 (Salsa, learn, methods)	0.6	0	-0.6
20	2	速読法,効果 (Speed reading method, effects)	0.5	0.1	-0.4
24	10	テーピング,方法 (Taping, method)	0.7	0.3	-0.4
31	10	資格試験,情報処理,IT (Qualifying exam, information processing, IT)	0.4	0.1	-0.3
38	10	加速器,医療,治療 (Accelerator, medical treatment, treatment)	0.6	0.3	-0.3
58	0	信越本線,碓氷峠,方法 (Shinetsu main line, Usui Pass, method)	0.4	0.1	-0.3
10	10	オーロラ,条件,観測 (Aurora, conditions, observation)	0.7	0.5	-0.2
34	10	キューブリック,映画,感想 (Kubrick, film, impression)	0.4	0.2	-0.2
49	10	ポリフェノール,種類,効果 (Polyphenol, type, effect)	0.5	0.3	-0.2
14	10	夢,将来,努力 (Dreams, future, effort)	0.1	0	-0.1
18	10	ロープワーク,結び方 (Rope work, knots)	0.8	0.7	-0.1
19	10	梅,名所,東京 (Plum tree, place of interes)	0.2	0.1	-0.1
36	1	パイプオルガン,コンサートホール,住所 (Pipe organ, concert hall, address)	0.2	0.1	-0.1
41	10	印象派,モネ,美術館 (Impressionist, Monet, art museum)	0.5	0.4	-0.1
42	10	イースター,復活祭,キリスト (Easter, Christ)	1	0.9	-0.1
48	10	アントシアニン,ブルーベリー,視力 (Anthocyanin, blueberry, eyesight)	0.8	0.7	-0.1
57	10	亀,寿命 (Turtle, lifespan)	0.1	0	-0.1
13	10	京都,寺,神社 (Kyoto, temple, shrine)	0.5	0.5	0
16	10	ゲノム,創薬,動向 (Genome, drug design, trend)	0.1	0.1	0
23	10	絶滅,哺乳類,危機 (Extinction, mammals, crisis)	0.3	0.3	0
28	10	著作権,デジタルコンテンツ,ネットワーク (Copyright, digital content, network)	0.5	0.5	0
29	10	スピーカー,評価,比較 (Speaker, evaluation, comparison)	0	0	0
39	10	宮崎駿,アニメーション,映画 (Miyazaki Hayao, animation, film)	0.1	0.1	0
53	10	自動車,将来像,日本 (Automobile, future image, Japan)	0	0	0
56	1	変分法,入門 (Calculus of variations, introduction)	0	0	0
61	0	Tommyfebruary,川瀬智子,TheBrilliantGreen (Tommy February, Kawase Tomoko, The Brilliant Green)	0.1	0.1	0
62	7	柴犬,日本犬,特徴 (Shiba inu, Japanese dog, characteristic)	0.5	0.5	0
12	10	正月,雑煮,地方 (New Years, ozoni soup, locality)	0.8	0.9	0.1
22	10	株式投資,インターネット,入門 (Stock investment, internet, introduction)	0	0.1	0.1
27	3	宮部みゆき,書評,レビュー (Miyabe Miyuki, book review, review)	0.1	0.2	0.1
30	2	アカデミー賞,受賞者,歴代 (Academy Award, recipient, successive generation)	0	0.1	0.1
32	2	憲法第九条,解釈,意見 (Article 9 of the Constitution, interpret, opinion)	0.6	0.7	0.1
33	9	石川県,特産品,お土産 (Ishikawa Prefecture, local product, souvenir)	0.2	0.3	0.1
35	9	三国志,ゲーム,題材 (Sanguozhi (The Three Kingdom), game, theme)	0.4	0.5	0.1
46	9	天然酵母パン,店,場所 (Natural yeast bread, shop, location)	0.2	0.3	0.1
59	0	Nゲージ,NOゲージ,意味 (N gauge, HO gauge, meaning)	0	0.1	0.1
63	10	グレートバリアリーフ,オーストラリア,旅行 (Great Barrier Reef, Australia, travel)	0.1	0.2	0.1
11	10	遣唐使,習慣,文化 (Japanese envoy to Tang Dynasty China, customs, culture)	0.2	0.4	0.2
17	10	野球,ベースボール,比較 ("Yakyu" (Japanese baseball), American baseball,	0	0.2	0.2
43	10	シフォンケーキ,作り方,菓子 (Chiffon cake, directions, cake)	0.6	0.8	0.2
44	7	アロマセラピー,アロマオイル,アロマキャンドル (Aromatherapy, aroma oil, aroma candle)	0	0.2	0.2
47	3	カプサイシン,とうがらし,効能 (Capsaicin, capsicum, effect)	0.4	0.6	0.2
52	10	湖,水質,透明度 (Lake, water quality, clarity)	0.2	0.4	0.2
15	10	オゾン層,オゾンホール,人体 (Ozone, ozone hole, human body)	0.3	0.6	0.3
40	7	本上まなみ,主演作品,女優 (Honjo Manami, works starred in, actress)	0	0.3	0.3
60	0	世界樹,北欧神話,名前 (Yggdrasil the world tree, Norse mythology, name)	0.4	0.8	0.4
37	10	バイク,ツーリング,レポート (Motorbike, touring, report)	0	0.6	0.6