NTCIR-5 WEB Navi-2 Experiments at Osaka Kyoiku University — Page, anchor and title indexing, and in-link count, inter page and inter site link analyses—

Takashi SATO* Hitoshi NAKAKUBO**
* Osaka Kyoiku University
4-698-1 Asahigaoka, Kashiwara, Osaka, Japan
sato@cc.osaka-kyoiku.ac.jp
** Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, Japan
hitosh-n@is.naist.jp

Abstract

This paper describes experimental results of WEB Navigational Retrieval Subtask 2 (WEB Navi-2). We made three gram-based indices, namely indices for text in whole page, text in title tag and text in anchor tag. Since gram-based indices are able to index all strings in target text, words that are not found in dictionaries are also indexed essentially. We used words in TITLE tag of search topics as queries. We did three kinds of link analyses, that is, in-link count and inter site and inter page link analysis. We merged score from word search for three indices and score from link analyses variously. We found that anchor text analysis was most effective for WEB Navi-2, and that it is necessary to devise merging of page and/or title score to anchor score.

Keywords: *gram-based index, page scoring, link analysis, score merge, NTCIR*

1 Introduction

The Web retrieval is daily activity for many people in the world, and the retrieval that is high-speed and accurate is expected. Moreover, the indexing and the link structure analyses for retrieval should be efficient because the Web space is not only huge but also it keeps growing. WEB Navigational Retrieval Subtask 2 is to search Known Item[1]. Since the number of pages (documents) that hit to each topic is few, a scoring and ranking technique different from conventional corpus retrieval is required. It is also the purpose of this task to experiment on the retrieval that positively uses the link structure of Web. We report on the result of score merge of the word retrievals and the link analyses we have done.

2 Indexing and word search

We made three gram based indices[2,3,4], namely indices for text in whole web page, text in title tag and text in anchor tag. Since gram based indices are able to index all strings in target text, words that are not found in dictionaries are also indexed essentially. The index for page, title, and anchor were divided into 222 pieces respectively according to the division of *sitelist* and *doclist*. We used words in TITLE tag of search topics as queries for the word retrieval. Three indices were retrieved with the same queries.

2.1 Page and title scoring

Score of page was calculated as follows. (1) The occurrence count *tf*, which is regularized by the length of page, for each query word was obtained. (2) From *tf* of each word, the value calculated by pseudo-probability function[1] ($f(tf)=(tf/(k+tf)$) was obtained. (3) The values obtained in (2) were summed over the word that composed each query weighting by *idf*s, which is log of the reciprocal of

[1] We call $f(x)=x/(k+x)$ as pseudo-probability function because $f(x)$ maps non-negative x to [0,1). Here, $k(\geqq 0)$ is constant value, and we set it by distribution of x.

the ratio of the number of documents in which the words are contained to the total page number. The title score was calculated similar way to page, considering length factor as the ratio of the title text length to the retrieval word length. As other scoring for page, we prepared the one in which kind of retrieval word was emphasized, that is, for each topics (kind of retrieval word in page / number of retrieval words for query) was multiplied.

2. 2 Anchor scoring

We calculated anchor score in similar way for page, and adopted the maximum value as a score of page the anchor ahead at the time of submitting runs. However, we thought that length of anchor text against retrieval word and the numbers of anchored are more important than similarity of anchor text itself, and we recalculated anchor score. Concretely, we gave value (retrieval word length / length of the anchor text) to each anchor, and we applied the pseudo-probability function to *va* which is sum of the value over linked anchor for each pages.

3 Link Analyses

We did three kinds of link analyses, that is, *in-link* count and *inter site* and *inter page* link analyses.

3.1 In-link count

The log of the number of *in-link*s to page *i* from another site page +1 was assumed to be l_i. Here, we defined site as the site information part of the first half of seven digits of *docid*s. Applying pseudo-probability function to l_i, we get the score of page *i*.

3. 2 Inter site link analysis

We extracted the site information part of the first half of seven digits of *docid* from *linklist*, which shows link structure between pages, and obtained the link structure between sites. We calculated PageRank[5] of the site link matrix. We took log of each elements of eigenvector, which belongs to the maximum eigenvalue, divided by the minimum element. We assume that the size of a site is the number of pages in the site. We applied the pseudo-probability function to ratio *vs* that was above-mentioned PageRank value and site size, and obtained the inter site link score. We assigned this score to pages in the site.

3. 3 Inter page link analysis

We calculated eigenvector from the in-link matrix among pages by a usual PageRank algorithm. We applied pseudo-probability function to log of corresponding elements of eigenvector divided by minimum elements, and we obtained the inter page link score of pages.

4 Score Merging and Ranking

As for the word scores, we calculated weighted sum of page, title, and anchor score. As for the link scores, we used at most one of them. In merging between the word and the link score, we thought the addition of the link score of pages with no relation to retrieval words to be harmful. We added the link score only to pages that had the word score. **Table 1** shows the combination of score merge we tried.

The sign "+"s in the table show that scores written before and behind are added. Moreover, *pg* stands for page, *ttl* stands for title, *anc* stands for anchor score respectively, and /2 and /4 show that merge is weighted 1/2 and 1/4 respectively. *Inlnk* stands for *in-link* count, *stlink* stands for *inter site link*, and *fllink* stands for *inter page (full) link* analysis score respectively.

Pg-k stands for scoring in which kinds of retrieval word was emphasized described in **2.1**.

5 Results

The evaluation results we obtained using the tool offered by NTCIR WEB TASK organizer are shown in **Table 2**. First columns of these tables, that is, from OKSAT-00 to OKSAT-44 indicate score combination of **Table 1**. From second columns, value and order pairs of *treceval R-precision, DCG* of three sets of gains, and *WRR* of two sets of parameter follow. Cut-off ranks for measuring *DCG*s and *WRR*s are 10 (default). OKSAT-00 corresponds to OKSAT-WEB-F-00 of submitted run id, and so on. It should be noted that anchor is re-calculated after we submitted runs as described in **2.2**.

Table 1. Combination of score merge

OKSAT-00	pg-k
OKSAT-01	pg
OKSAT-02	pg+anc
OKSAT-03	pg/2+anc
OKSAT-04	pg-k+anc
OKSAT-05	pg-k/2+anc
OKSAT-06	pg+anc+inlnk
OKSAT-07	pg/2+anc+inlnk
OKSAT-08	pg+anc+stlnk
OKSAT-09	pg/2+anc+stlnk
OKSAT-10	(pg+fllnk)+anc
OKSAT-11	(pg+fllnk)/2+anc
OKSAT-21	pg+ttl+anc+stlnk
OKSAT-22	pg/2+ttl+anc+stlnk
OKSAT-23	(pg+ttl)/2+(anc+stlnk)
OKSAT-24	ttl
OKSAT-25	pg+ttl
OKSAT-31	anc
OKSAT-32	anc+stlnk
OKSAT-33	anc+fllnk
OKSAT-34	anc+inlink
OKSAT-41	(anc+stlnk),(pg+ttl)/2
OKSAT-42	(anc+stlnk),(pg+ttl)/4
OKSAT-43	(anc+inlnk),(pg+ttl)/4
OKSAT-44	(anc+fllnk),(pg+ttl)/4

We thought that the retrieval of page text is most basic, we fist ranked pages by their score only (OKSAT-00, 01). Next, we expected the effect of anchor and the link analysis, and added them to page score (OKSAT-02, 03, 04, 05, 06, 07, 08, 09, 10, 11). After submitting runs, we expected the effect of title tag, so we made title index and got runs (OKSAT-21, 22, 23, 24, 25).

As experimenting using evaluation tools from organizer, we noticed that page and title score do not work effectively. So we removed page and title score from ranking (OKSAT-31, 32, 33, 34). Consequently, we observed improvement. We realized simple addition of page and title score is harmful. By topic-by-topic investigation, we observe topics for which these simple addition is effective. That is, if we devise the usage of page and title score, ranking may be improved.

Then we changed the way of score merging for OKSAT-41, 42, 43, 44. That is, we regarded anchor and link score was first group and page and title score was second group, then we extracted top 70 of first group and top 30 of second group instead of simple addition of these groups. More improvement was observed.

As the entire tendency, evaluation results are better in the order of (A) page+title (OKSAT-00, 01, 24, 25), (B) simple addition of anchor+link group to page+title group, (C) anchor+link (OKSAT-31, 32, 33, 34), (D) Top rank extraction from anchor+link group and page+title group. However, the difference between (C) and (D) is not large, and the order of evaluations are overlapped.

Link analyses had certain effects. As for the cost analysis of a *full link*, the calculation cost was high, however, the effects were not so larger than other link analyses. This time we grouped pages by site, we think we have to investigate grouping method[6].

Evaluation results using relevance A and AB were same tendency.

6 Discussions

6.1 Analyses of runs

When the number of pages that matches to the retrieval target is few as this WEB Navigational Retrieval Subtask 2, it is difficult to answer relevant pages by page text only. We got relatively good evaluation results by anchor text analyses. It is effective in scoring anchor text that we consider its length. Score of page and title retrieval is effective, however, simple addition of these score to anchor score has the opposite effect.

6.2 Failure examples topic-by-topic

We show some failure examples. TITLE of topics and failure reasons follow topic-id, because we use TITLE tag of topics only.

Topic#1006: {ANA, **オンラインチケット**} : Word "**オンラインチケット**" is rarer than word "ANA", then pages that have "**オンラインチケット**" are over scored.

Topic#1010: {bunkamura **ミュージアム**}: We retrieved this long word only. We should retrieve "bunkamura " and "**ミュージアム**" also.

Topic#1013: {ExCite, **英和**}: Other than page

score is low, while Page score is high.

Topic#1014: {FP, 資格}: Both two words are too much popular.

7 Conclusions

We describe our experimental results of WEB Navigational Retrieval Subtask 2 We made three gram based indices, namely indices for text in whole page, text in title tag and text in anchor tag. We did three kinds of link analyses, that is, in- link count and inter site and inter page link analyses. We merged score from word search for three indices and score from link analyses variously. We found that anchor text analysis was most effective for WEB Navi-2, and that it is necessary to devise merging of page and/or title score to anchor score.

References

[1] NTCIR-WEB, (http://research.nii.ac.jp/ntcweb/).

[2] Sato, T., Fast full text search with free word using TS-file, *Proc. 19th ACM SIGIR Conf.*, p.342 (1996).

[3] Sato, T., Fast full text retrieval using gram based tree structure, *Proc. ICCPOL '97*, Vol.~2, pp. 572--577 (1997).

[4] Sato, T. *et al.*, Gram based full text search system and its application, *IPSJ SIG Notes*, 98-DBS-114-2 (1998).

[5] Brin, S. and Page, L., The anatomy of a large-scale hypertextual web search engine, *In Proceedings of the 7th International World Wide Web Conference (WWW7)*, pp.107-117, 1998.

[6] Nakakubo, H. and Sato, T., Static and dynamic ranking by web page grouping, *Proceedings of Data Engineering Workshop 2005 (DEWS2005)*, 5C-i6, 2005.

Table 2. Evaluation Results (Relevance A)

	trec		dcg.3-3-0		dcg.3-2-0		dcg.3-0-0		wrr.1-1-0		wrr.1-0-0	
	R-prec.	ord	val	ord	val	ord	val	ord	val	ord	val	ord
OKSAT-00	0.0611	23	0.8634	23	0.6581	23	0.2475	24	0.0928	24	0.0439	24
OKSAT-01	0.0609	24	0.8156	24	0.6247	24	0.2430	25	0.0885	25	0.0432	25
OKSAT-02	0.1216	20	1.7391	20	1.5385	20	1.1373	20	0.3092	19	0.2518	19
OKSAT-03	0.1511	13	2.0464	13	1.8565	13	1.4768	13	0.3794	12	0.3165	13
OKSAT-04	0.1332	16	1.8270	17	1.6025	19	1.1533	19	0.3137	18	0.2564	18
OKSAT-05	0.1590	09	2.1395	10	1.9444	10	1.5543	10	0.3920	10	0.3277	11
OKSAT-06	0.1326	18	1.8387	16	1.6301	16	1.2130	16	0.3356	16	0.2696	16
OKSAT-07	0.1547	12	2.1058	11	1.9092	11	1.5160	11	0.3954	09	0.3286	09
OKSAT-08	0.1266	19	1.8058	19	1.6031	18	1.1977	18	0.3206	17	0.2633	17
OKSAT-09	0.1550	11	2.0937	12	1.8980	12	1.5065	12	0.3881	11	0.3279	10
OKSAT-10	0.1129	21	1.5863	21	1.4207	21	1.0895	21	0.2702	21	0.2265	21
OKSAT-11	0.1411	15	1.9702	15	1.7939	15	1.4414	15	0.3645	14	0.3056	14
OKSAT-21	0.1328	17	1.8072	18	1.6047	17	1.1995	17	0.2997	20	0.2477	20
OKSAT-22	0.1413	14	2.0360	14	1.8415	14	1.4525	14	0.3433	15	0.2906	15
OKSAT-23	0.1553	10	2.1588	09	1.9611	09	1.5656	09	0.3782	13	0.3207	12
OKSAT-24	0.0565	25	0.6821	25	0.5871	25	0.3971	23	0.1015	23	0.0719	22
OKSAT-25	0.0731	22	2.3559	22	0.8467	22	0.4053	22	0.1265	22	0.0638	23
OKSAT-31	0.1891	06	2.4545	08	2.1879	08	1.8517	08	0.4959	07	0.4447	06
OKSAT-32	0.1915	04	2.2772	06	2.2743	06	1.9138	06	0.5108	05	0.4593	01
OKSAT-33	0.1867	07	2.5082	07	2.2284	07	1.8771	07	0.4958	08	0.4414	07
OKSAT-34	0.1996	02	1.0674	05	2.3217	04	1.9486	04	0.5110	04	0.4580	04
OKSAT-41	0.1795	08	2.5630	02	2.3627	02	1.9621	02	0.5113	02	0.4593	01
OKSAT-42	0.1977	03	2.5630	02	2.3627	02	1.9621	02	0.5113	02	0.4593	01
OKSAT-43	0.2035	01	2.6167	01	2.4101	01	1.9969	01	0.5115	01	0.4580	04
OKSAT-44	0.1903	05	2.5126	04	2.3169	05	1.9255	05	0.4964	06	0.4414	07