Static and Dynamic Scoring by Web Page Grouping

Hitoshi NAKAKUBO* and Takashi SATO**
** Course of Information Science, Graduate School of Education, Osaka Kyoiku University*
*** Information Processing Center, Osaka Kyoiku University*
** nakaku@ss.osaka-kyoiku.ac.jp ** sato@cc.osaka-kyoiku.ac.jp*

Abstract

Web Search System exists to retrieve necessary information on the WWW space. However, these are not accuracy enough. Then, we propose the technique for using Web Page Grouping together with the link structure analysis, and aim at the improvement in accuracy. Our proposal is composed of four techniques. The first technique is Web Page Grouping to enhance it related to adjacent the link structure. The second technique is static scoring by link structure analysis using Web Page Grouping. The third technique is dynamic scoring which tempers the link structure analysis of the retrieval result set with Web Page Grouping. And, the fourth technique is a ranking technique for annexing a static score and a dynamic score. This paper describes these techniques, and reports on the experiment results.

1. Introduction

Information, which exists on the WWW space, is large-scale. The quality of information is various. Therefore, the extraction of information, which the Internet user needs, is difficult. Web Search System exists to assist this difficult work. However, when Web Search System is used, necessary information cannot be effectively extracted. Then, the accuracy improvement of Web Search System is needed.

General Web Search System is retrieved based on the query that the Internet user input. Web Search System extracts agreement web page group by retrieving each web page text. However, there is a limit in the accuracy, which can be obtained by a simple full-text search technique for using only query and web page text. Then, the technique for effectively using the HTML document structure and the link structure, which is peculiar information to web pages, is examined.

PageRank algorithm[1][2] and HITS algorithm[3] are enumerated as typical techniques using a link structure analysis. These are techniques for deciding the ranking based on the adjacent relations between each web pages, and relations to web pages which does not exist in adjacent relations are recurrently solved. However, a link structure does not necessarily exist between relating web pages on the WWW space of the reality. For instance, when the web page to have permitted the link act only to the web site top page exists, it is thought that an appropriate result can be obtained in the ranking decided based on the adjacent relation.

We propose a technique, which enhancing adjacent relations of a link structure by making web page group with similar information a group, to reduce this problem. Moreover, to prevent the accuracy decrease because of enhancing related to adjacent a link structure. We propose Dynamic Scoring technique.

We describe related works in Chapter 2, and our proposal technique in Chapter 3, describe experiments in Chapter 4 and 5, consider in Chapter 6, and describe conclusions in Chapter 7.

2. Related works

2.1. PageRank algorithm

PageRank algorithm is an index, which shows the importance degree on web pages. This technique defines link act, "Act of recommending linking ahead". The score obtained by this technique clearly shows the level from which each Web page is referred. However, link act is not necessarily possible compared with web pages, which wants to be recommended. It is thought that the problem of this technique.

2.2. HITS algorithm

The HITS algorithm is a technique for the community extraction. This defines Authority and Hub as an index, which shows the importance degree on the web page, and defines these relations of two, "Good quality Authority is linked by two or more, good quality Hub and good quality

Hub has been linked with two or more good quality Authority". The Authority score clearly shows web pages useful group as information, and the Hub score shows web pages useful group clearly as links. However, this technique has a problem that an appropriate community cannot be necessarily extracted anytime because the full-text search result is not used to extract the community candidate.

3. Proposal

We indicate our proposal technique[4] as follows.

3.1. Web Page Grouping

We consider web page set with similar information is treated as a group to enhance the adjacent relation of the link structure. We defined web page set with similar information, "The web page set of the similar information and being made by the same author". We used the directory structure to process making to groups. In this case, "Same manufacturer" can be judged by deciding the web site district switching off from the URL character string. Moreover, "web page set thought that it is similar information" can be judged by deciding web page set included from the URL character string in the same directory. Web pages are made to groups from these judgments. Enhancing related to adjacent deleting a link structure between web pages in a group, and substituting it for a link structure between groups achieve a link structure.

3.2. Static Scoring

Static scoring applies the link structure analysis to the link structure after Grouping all Web page on the WWW space. We think that we can reduce the problem in the PageRank algorithm by Grouping.

3.3. Dynamic Scoring

Dynamic scoring applies link structure analysis to a link structure of the full-text search result set. In that case, two scores are calculated. The first is a score concerning the link structure composed between Web pages in the full-text search result set, and it shows a clear level to be referred. Another is a score concerning the link structure in the full-text search result set after Grouping, and it shows the score in which the adjacent relation of the link structure is enhanced. We think that we can reduce the problem in the HITS algorithm because we calculate the score based on the web page or the group in the full-text search result set.

3.4. Ranking

The annexation score is calculated by annexing static and dynamic scores to the full-text search score, and the rank is decided. The annexation score is calculated by adding after each score is regularized, and the weighting factor is multiplied respectively.

4. Experiment environment

The environment used to experiment is shown.
Full-text search system: Variable-length gram base index[5] was used.
Retrieval target: Test collection NW100G-01 which had been offered in NTCIR-4 Web Task[6][7] was used.
Retrieval query: 77 queries were used from retrieval queries, which had been offered in NTCIR-4 Web Task B.
Evaluation method: WRR[8][9] and 11-point Recall-Precision curve were used.

5. Experiment result

5.1. Web Page Grouping and Static Scoring

Table 1 shows the result of Grouping to the retrieval target. Uneven to the number of web pages included in each group is known. The difference of the number of web pages is thought that it influences the number of links, and thought to have to reexamine Grouping.

Table 2 shows the comparison result before and after Grouping. The maximum value before Grouping is higher than after Grouping. Moreover, the minimum value after Grouping is higher than before Grouping. In addition, the range of the score is greatly different before and after Grouping. We think that the influence given to the rank is large the score before Grouping, and is small the score after Grouping.

Table 3 shows the ratio of the retrieval query, which extracts relevant documents about of each before and after Grouping. In figure, “Both" shows the query, which could be extracted by both before and after Grouping. Whether it is an evaluation of “Both” that either is better before or after Grouping is shown. In the ratio of extractive retrieval query, before Grouping is 61%, and after Grouping is 13%. This means retrieval query extractive only by after Grouping existed by 12%.

Therefore, we think that each score is mutually opposite before and after Grouping. Thus, we think that scoring with both characters in annexing two scores is possible. The expression which calculates *ScoreS(p)* in document *p* as a static score is shown below. It is shown respectively that *Retrieval* is a full-text search score, and

StaticN is a Static Score before Grouping, and *StaticG* is a Static Score after Grouping, and *W* is the weighting factor.

$$ScoreS(p) = Wr \cdot Retrieval(p) + Wsn \cdot StaticN(p) + Wsg \cdot StaticG(p)$$

Figure 1 shows the Static Scoring evaluation result. The accuracy improvement can be confirmed from the result by (*Wr*, *Wsn*, *Wsg*) = (1, 1, 1).

5.2. Dynamic Scoring and Ranking

Table 4 shows the Grouping result in the full-text search result set. The score became a tendency to look like static scoring.

Table 3 shows the ratio of the retrieval query, which extracts an appropriate document about of each before and after Grouping. In the ratio of extractive retrieval query, before Grouping is 32%, and after Grouping is 31%. This means relevant documents that of each is different is extractive as the same. Thus, we think that we can extract a lot of relevant documents by annexing two Dynamic Scores.

We think a final score is calculated by annexing all of each score because each score of the proposal technique is a score with the feature respectively. The expression which calculates *Score(p)* in document p as the final score is shown below. It is shown respectively that *DynamicN* is a Dynamic Score before Grouping, and *DynamicG* is a Dynamic Score after Grouping.

$$Score(p) = Wr \cdot \mathrm{Re}\,trieval(p) + Wsn \cdot StaticN(p) + Wsg \cdot StaticG(p) + Wdn \cdot DynamicN(p) + Wdg \cdot DynamicG(p)$$

Figure 2 shows the final scoring evaluation result. It can be confirmed that it is the best performance for (*Wr*, *Wsn*, *Wsg*, *Wdn*, *Wdg*) = (2, 1, 2, 0, 0) from the result. Moreover, when a Dynamic Score is annexed, it can be confirmed that the performance has decreased.

6. Consideration

The WRR at rank 100 evaluation results of the best result in the proposal technique and other techniques are compared. As a result, the evaluation improvement of about 200% was confirmed compared with the case to use only the full-text search score. Moreover, the evaluation improvement of about 10% was confirmed compared with the case to use only the PageRank algorithm. However, when all scores were annexed, the result of evaluated decreasing was confirmed. In addition, the result of evaluated greatly decreasing was confirmed when a Dynamic Score was annexed. We think that the cause of the evaluation decrease is accuracy shortage of Dynamic Scoring.

Then, we investigated the evaluation when each Dynamic Score was annexed to the full-text search score before and after Grouping. As a result, it was confirmed that the evaluation when a Dynamic Score was annexed had decreased compared with the evaluation only according to the full-text search score. It seems that the evaluation when a Dynamic Score is annexed after the full-text search score and Grouping are applied is especially bad, and the evaluation when all scores are annexed is negatively affected. Moreover, the low degree of the accuracy of Grouping is thought as one of causes with a bad evaluation.

7. Conclusion

In this paper, we proposed to enhance it by making web page set with similar information a group related to adjacent the link structure. And, we proposed the ranking technique for applying the link structure analysis statically and dynamically. Moreover, we experimented on the proposal technique.

As a result, we confirmed the accuracy improvement by the proposal technique, and we confirmed the technique of making the group had to be reexamined. We reexamine the technique of making to the group, and will verify the expression of the annexation score calculation in the future.

References

[1] S. Brin and L. Page, "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.

[2] L. Page, "The PageRank Citation Ranking: Bringing Order to the Web," http://google.stanford.edu/~backrub/pageranksub.ps, 1998.

[3] J. M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *Journal of the ACM*, vol.46, no.5, pp.604-632, 1999.

[4] Hitoshi NAKAKUBO, and Takashi SATO, "Ranking Method Using Link Structure Analysis in Web Retrieval," DBWS2004, *Technical Report of IEICE*, Vol.104 No.177, no. DE2004-65, pp99-103, Jul. 2004.

[5] T. Sato, T. Satomoto, and K. Han, "NTCIR-3 PAT Experiments at Osaka Kyoiku University," In *Working Notes of the 3rd NTCIR Workshop Meeting Part III: Patent Retrieval Task*, pp.21-24, Tokyo, Japan, 2002.

[6] K. Eguchi, K. Oyama, A. Aizawa, and H. Ishikawa, "Overview of WEB Task at the Fourth NTCIR Workshop," In *Working Notes of the Fourth NTCIR Workshop Meeting*, pp.ov1-ov2, Tokyo, Japan, 2004.

[7] K. Eguchi, K. Oyama, A. Aizawa, and H. Ishikawa, "Overview of the Information Retrieval Task at NTCIR-4 WEB," In *Working Notes of the Fourth NTCIR Workshop Meeting*, pp.ov3-ov15, Tokyo, Japan, 2004.

[8] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama, "Overview of the Web Retrieval Task at the Third NTCIR Workshop," *NII Technical Report*, NII-2003-002E, 2003.

[9] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama, "Evaluation Methods for Web Retrieval Tasks Considering Hyperlink Structure," *IEICE Transactions on Information and Systems*, vol. E86-D, No.9, pp.1804-1813, 2003.

Table 1. Result of Grouping

Minimum	1
Maximum	30,446
Average	5
Median	1

Table 2. Comparison Result before and after Grouping in Static Scoring

	Before	After
Node Total	23,670,000	4,500,000
Link Total	79,700,000	18,140,000
Score Maximum	2.6126E-04	4.1985E-07
Score Minimum	7.3143E-09	3.3442E-08
Score Average	4.2231E-08	2.2230E-07
Score Median	8.3860E-09	2.2612E-07

Table 3. Ratio of Retrieval Query

		Static	Dynamic
Before		49%	19%
Both	Before	12%	14%
	After	1%	13%
After		12%	17%
None		26%	37%

Table 4. Comparison Result before and after Grouping in Dynamic Scoring

	Before	After
Node Total	192,500	124,041
Link Total	95,848	120,292
Score Maximum	4.8634E-01	5.6874E-02
Score Minimum	6.8460E-05	7.6747E-05
Score Average	4.0000E-04	6.3694E-04
Score Median	7.0123E-05	5.1010E-04

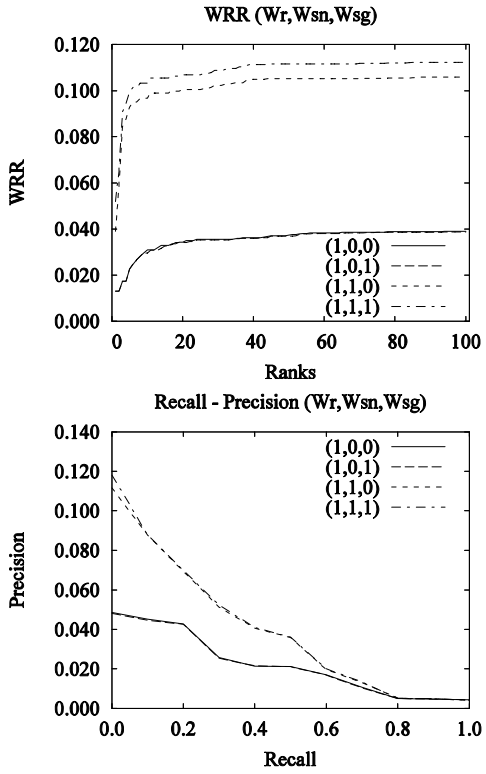

Figure 1. Static Scoring Evaluation Result

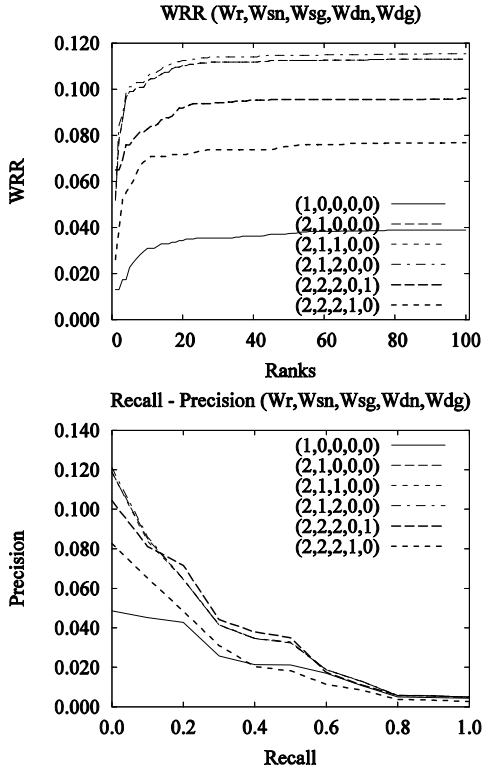

Figure 2. Final Scoring Evaluation Result