Discriminative Power and Retrieval Effectiveness of Phrasal Indexing Terms

Sumio Fujita

Justsystem corporation
Brainspark, Tokushima, Japan
Email: Sumio_Fujita@justsystem.co.jp

Abstract

In spite of long controversy, effectiveness of phrasal indexing is not yet clear.
Recently, correlation between query length and effect of phrasal indexing is reported.
In this paper, terms extracted from the topic set of the NACSIS test collection 1 are analyzed utilizing statistic tools in order to show distribution characteristics of single word/phrasal terms with regard to relevant/non-relevant documents. Phrasal terms are found to be very good discriminators in general but not all of them are effective as supplemental phrasal terms. A distinction of informative / neutral / destructive phrasal terms is introduced. Retrieval effectiveness is examined utilizing query weight ratio of these three categories of phrasal terms.

Introduction

Longer queries are not necessarily better than shorter queries in view of retrieval effectiveness, since longer queries may contain so-called noisy terms that hurt the performance.
Given relevance judgements, we can say which terms are noisy and which are not with regard to a certain topic description and a test collection.
We can confirm that a term is good to discriminate subject concepts if relevant documents contain such terms and non-relevant documents do not contain them and that a term is noisy if the situation is the opposite.
The problem here is that not only noisy terms but also good terms can harm the performance in some cases where term weighting is not adequate or terms are redundant.
One example of such cases is complex terms like supplemental phrases or overlap bigrams which violate term independence assumption.
Phrasal terms are utilized either as replacement of single words or as supplemental units for single words, but according to our experience, phrasal terms as replacement of single words do not perform well. Supplemental phrasal terms works better in spite of the violation of term independence assumption.
Recent studies uncovered the correlation between phrase effectiveness and query length(Fujita, 2000).
In this paper, we will see the problem of effectiveness of phrasal terms from two different viewpoints utilizing a large test collection for Japanese text retrieval and statistical tools.
NACSIS test collection 1(NTCIR, 1999), which consists of a collection of abstracts of scientific papers (330,000 records, 590MB in text), two sets of topic description (30 topics for training and 53 topics for evaluation) and relevance judgement, provides us of a good opportunity for this purpose.
Topic description of NACSIS test collection 1 contains four different fields, just like early versions of TREC topics, as follows:
<title> fields consist of one (typically simple) noun phrase.
<description> fields consist of one (typically simple) sentence.
<narrative> fields consist of 3 to 12 sentences and contain detailed explanation of the topic, term definition, background knowledge, purpose of the search, preference in text types, criteria of relevance judgement and so on.
<concepts> fields consist of lists of keywords corresponding to principal concepts in the information need.
Combining these four fields, different length of query sets for the same topics are prepared.

Topic field used	Avg.Prec1 (Single words only)	Avg.Prec2 (Single words & Phrases)	Avg.Prec2 – Avg.Prec1	Avg. number of total terms	Avg. number of phrasal terms
<description>	0.3143	0.2846	-0.0297	8.8	1.9
<title>	0.2555	0.2265	-0.029	4.1	1.0
<title>,<description>	0.3334	0.3079	-0.0255	9.2	2.1
<title>,<narrative>	0.3095	0.3001	-0.0094	45.0	10.3
<narrative>	0.2985	0.2895	-0.009	44.7	10.2
<description>,<narrative>	0.3161	0.3163	0.0002	46.4	10.8
<title>,<description>,<narrative>	0.321	0.3233	0.0023	46.5	10.9
<description>,<concepts>	0.3672	0.3786	0.0114	25.4	5.2
<narrative>,<concepts>	0.364	0.3761	0.0121	57.0	12.5
<title>,<description>,<concepts>	0.379	0.3926	0.0136	25.5	5.3
<title>,<narrative>,<concepts>	0.3702	0.3844	0.0142	57.3	12.7
<description>,<narrative>,<concepts>	0.3681	0.3839	0.0158	58.4	13.1
<title>,<description>,<narrative>,<concepts>	0.371	0.3886	0.0176	58.4	13.1
<concepts>	0.3316	0.3504	0.0188	20.9	4.1
<title>,<concepts>	0.352	0.3711	0.0191	21.8	4.5

Table 1: Performance comparison using 15 different versions of queries combining 4 fields

1. Phrasal Indexing

For the baseline run experiments, we utilized the engine of Conceptbase Search 1.2, a commercial based search engine adopting vector space model approach.

1.1. Linguistic Phrases as Indexing Units for Japanese Text Retrieval

For automatic indexing of Japanese written text, once word boundary is detected by morphological analysis processing, word based approach normally adopted in English IR can be applied. Although computationally more expensive than in English, the accuracy of Japanese morphological analysis is quite high and sufficient for IR purpose.

Our approach consists of utilizing noun phrases extracted by linguistic processing as supplementary indexing terms in addition to single word terms contained in phrases. Phrases and constituent single word terms are treated in the same way, both as independent terms, where the frequency of each term is counted independently based on its occurrences.

Linguistic phrases are normally contiguous kanji or katakana word sequences and internal phrase structures are ignored.

1.2. Query Length and Effectiveness of Phrasal Indexing

Among evaluation experiments of the NTCIR1 workshop, correlation between query length and the effect of phrasal indexing is reported in (Fujita, 1999).

NTCIR topic description consists of four fields namely <title>, <description>, <narrative> and <concepts> as shown in the previous chapter. The combination of these four fields makes 15 different versions of queries for each topic. These 15 different versions of queries for 53 topics are examined with phrasal terms and with only single word terms.

Table 1 shows the performance with 15 versions of queries, where we compared two types of indexing language in question i.e. single words vs. single words + supplemental phrases. Performance is indicated as non-interpolated average precision macro averaged for 53 topics.

Since this experiment is designed to clarify the effect of different length of queries, the following settings are chosen:

1) no pseudo feedback procedure is processed,
2) no down-weighting coefficient is applied for phrasal terms,
3) no field specific importance coefficient is applied.

Consequently, absolute performance is much worse than our best performing runs.

Out of 15 versions of query sets, 10 times phrasal indexing performs better than single word only indexing, and 5 times vice versa. This is exactly the situation described in literature that the effect of phrasal indexing is inconsistent and uncertain.

We found out that there is clear correlation between the difference of average precision and number of terms contained in the query. Pearson's correlation coefficient between Avg. prec2 - Avg.prec1 and average number of terms accounts for 0.57, while 0.52 between Avg. prec2 - Avg.prec1 and average number of phrases. Eliminating 8 query versions containing <concepts> field, correlation coefficients become 0.96 and 0.95 respectively.

<concepts> fields containing keywords that are essentially noun phrases, tend to favor phrasal indexing otherwise when using only one of the fields, single word runs perform better.

The situation is different when more than two fields are combined. Combining <title>, <description> and <narrative> fields, the supplemental phrasal run performs better than the single word run.

We can see that the length of query, which is number of features in the scoring function, is important factor as well as quality of phrasal terms extracted from topic description, in order to evaluate phrasal indexing.

Two aspects of characteristics of phrasal terms should be considered:

1) Are the phrasal terms good discriminator of subject domain?
2) Do the supplemental phrasal terms cause some undesirable influence to original word based queries?

In the chapter 2, phrasal terms extracted from the topic set of the NACSIS test collection 1 are examined from the viewpoints of their discriminative power. In the chapter 3, we will see another aspect of retrieval effectiveness.

Df	Term
98561	研究(research)
83016	結果(result)
69911	３
64675	報告(report)
63956	特性(characteristics)
61063	構造(structure)
58664	方法(method)
58410	システム(system)
56807	解析(analysis)
50246	影響(influence)
47620	評価(evaluation)
42130	利用(use)
41584	モデル(model)
41238	処理(process)
37567	時間(time)

Table 2: High document frequency single word terms

Df	Term
12817	有効 性(effectiveness)
6969	３ 次元(3-dimension)
5716	モデル 化(modeling)
5183	効率 的(efficient)
4659	光 ファイバー(optic fiber)
2648	利用 者(user)
1795	高齢 者(old people)
1661	有効 利用(effective use)
1347	遺伝 的 アルゴリズム (genetic algorithm)
1345	階層 的(hierarchy)
1038	ａ ｔｍ 網(ATM network)
860	グループ ウエア(groupware)
799	人工 知能(artificial intelligence)
777	データ 転送(data transmission)
672	分散 環境 (distributional environment)

Table 3: High document frequency phrasal terms

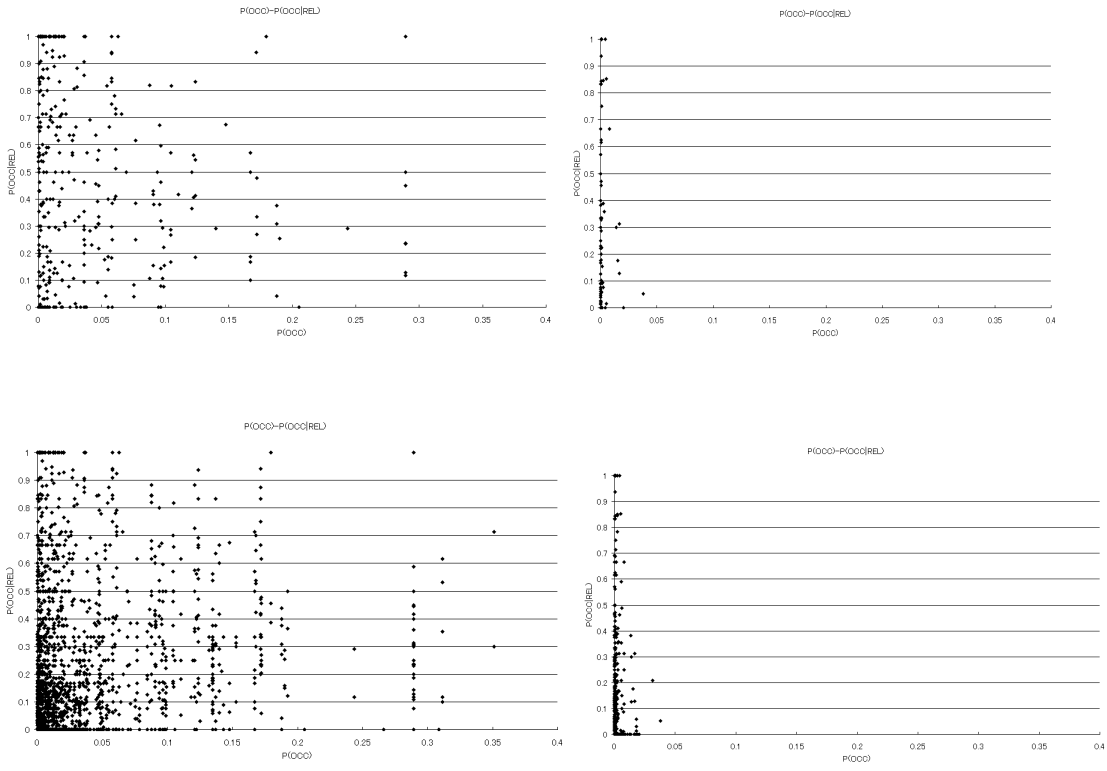

Figure 1: p(occ|rel) as function of p(occ)

Left above: short query single words, Right above: short query phrases

Left below: long query single words, Right below: long query phrases

2. NTCIR Data Analysis

Greiff presented an analysis of TREC data plotting each query terms in view of distributions in the whole document collection and in relevant document sets(Greiff, 1998) and Pickens et al. applied this analysis for statistical phrases(Pickens et al, 2000).

Adopting their plotting approach, we will try to clarify distribution characteristics of phrasal terms using mainly p(occ|rel) and p(occ) which are computed as document frequencies of the term in relevant documents /the whole collection respectively divided by each number of documents.

2.1. Occurrence in Relevant Documents and in Non-relevant Documents

Table 2 and Table 3 shows high document frequency terms extracted from the short query set of test topics.

A short query refers to a query constructed using only <description> field of topic description and a long query, all fields of topic description.

First, plotting of p(occ|non-rel) as function of p(occ) is not interesting since approximately the relation p(occ|non-rel)=p(occ) is observed. This is not surprising because number of relevant documents are generally very small and p(occ|non-rel) can be approximated by p(occ).

From Table 2 and Table 3, we can imagine that the distribution characteristics of phrasal terms are almost same as single words i.e. Zipfian distribution but document frequencies of phrasal terms are much smaller than single words.

It seems difficult to get clear intuition about term distribution characteristics from Figure 1, where p(occ|rel) is plotted as function of p(occ). The same p(occ) value for some frequent terms found in plots indicates multiple occurrences of a term in different queries.

As Greiff suggests, a different visualization is desirable for this graph.

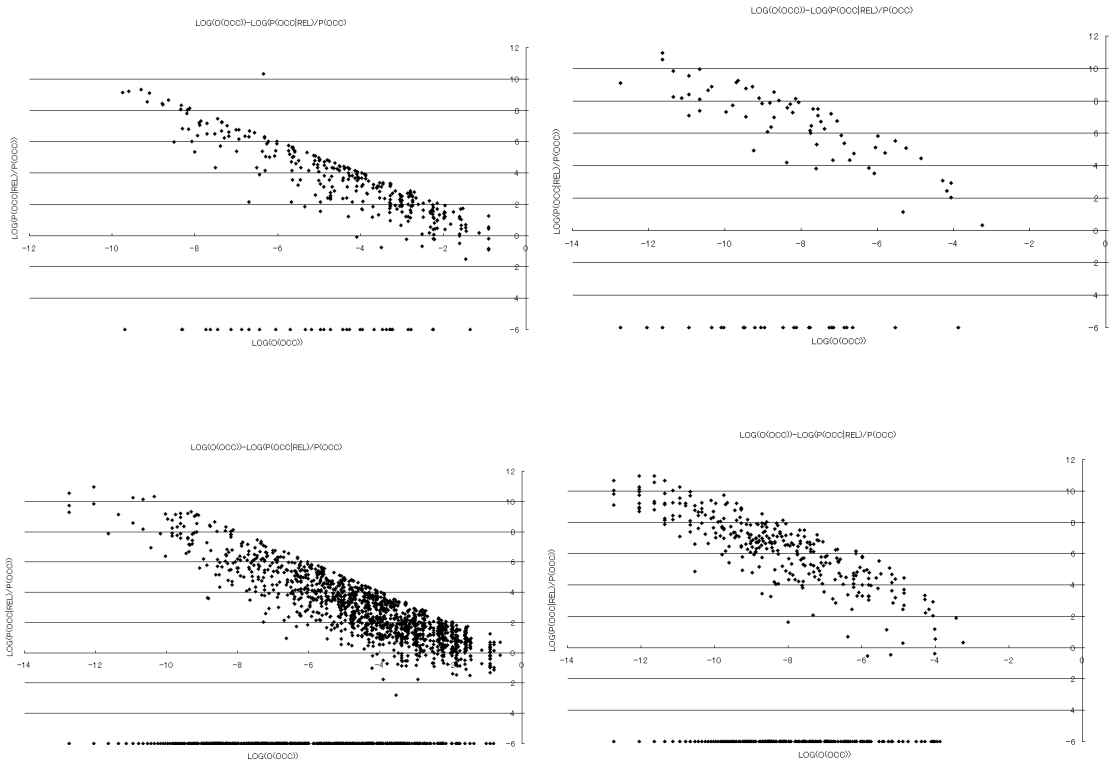

Figure 2: log(p(occ|rel)/p(occ)) as function of log(O(occ))

Left above: short query single words, Right above: short query phrases

Left below: long query single words, Right below: long query phrases

First p(occ) is replaced by log(O(occ))=log(p(occ)/1-p(occ)), since distribution of p(occ) is too skewed.

In Figure 1, if the dot representing a term located higher than the graph of p(occ)=p(occ|rel), the term can be a good discriminator and should contribute to retrieval performance given an adequate weighting scheme. On the other hands, the terms plotted lower than the graph of p(occ)=p(occ|rel) are by no means useful for retrieval performance irrespective of weighting scheme.

P(occ|rel) is replaced by log(p(occ|rel)/p(occ)) in order to illustrate this borderline. In the case of zero probability for p(occ|rel), -6 is assigned for log(p(occ|rel)/p(occ)).

This is equivalent to mutual information MI(occ;rel) in information theory as follows:

$$\log\left(\frac{p(occ \mid rel)}{p(occ)}\right) = \log\left(\frac{p(occ, rel)}{p(occ)\,p(rel)}\right) \quad (1)$$

Finally, Figure 2 illustrates distribution characteristics of terms much better than Figure 1.

The dots plotted above the y=0 line represent useful terms with respect to the query and

	Single words	Phrases	Single words + phrases
Short query	79.29%(291/367)	66.34%(67/101)	76.50%(358/468)
Long query	54.77%(1315/2401)	45.32%(315/695)	52.65%(1630/3096)

Table 4: Ratio of positive log(p(occ|rel)/p(occ)) for query terms

	Single words	Phrases	Single words + phrase
Short query	2.81	4.38	3.15
Long query	1.65	2.92	1.93

Table 5: Average of positive log(p(occ|rel)/p(occ)) value for query terms

relevance judgements.

As this shows, single words and phrases are very similar distribution characteristics but document frequencies for phrases are much lower. Average of log(O(occ)) is –5.22 for single words while –8.64 for phrases in long queries.

On the other hands, ratios of good terms, whose log(p(occ|rel)/p(occ)) is larger than 0, are shown in Table 4.

From this observation, we can see limited usefulness of phrasal terms with regards to relevance. The ratio of positive log(p(occ|rel)/p(occ)) is lower than single words. This explains poor performance of pre-coordinated longer phrase based indexing that utilizes phrases as replacements of single words.

Phrasal terms tend to have high value of log(p(occ|rel)/p(occ)), but this does not necessarily mean effectiveness of phrasal terms. As Figure 1 and Figure 2 illustrate, the terms with high log(p(occ|rel)/p(occ)) value tend to have low log(O(occ)) that means extremely lower document frequency so that they are not so useful because of such lower frequency.

2.2. Measures for Phrasal Term Effectiveness

Table 4 and Table 5 seem to support supplemental phrasal indexing, because fairly high ratio of positive log(p(occ|rel)/p(occ)) terms, and higher average value of log(p(occ|rel)/p(occ)) are observed. But for short queries, supplementing phrasal terms did not show any positive effect as we have seen in Table 1.

The following accounts are enumerated.

1) Over-weighted phrasal terms may cause topic deviation from concepts represented by single words to concepts represented by phrasal terms.

2) Supplemental phrasal terms are not always informative because their constituent single words are already indexed.

If the phrasal term AB has a high MI(AB,rel) value in contrast with MI(A,rel) and MI(B,rel), this is the case where phrasal terms are effective.

Consider a supplemental phrasal term as informative if and only if its MI(occ,rel) is positive value and is higher than the sum of MI(occ,rel) of constituent single words in view of the query and relevance judgements. A phrase "AB" is informative means that the occurrence of a phrase "AB" gives more information about relevance than occurrence of both single words "A" and "B".

Table 6 shows the number and the ratio of informative phrasal terms. –1 is assigned for MI(occ,rel) when p(occ|rel) is 0.

Giving different values (-3 and -6) for MI(occ|rel) when p(occ|rel)=0 did not change the results..

	{#phrasal terms| MI(AB,rel) > SUM(MI(A,rel), MI(B,rel))}	Positive MI(occ,rel) phrasal terms	Total phrasal terms
Short query	31(30.69%)	67(66.34%)	101
Long query	146(21.01%)	315(45.32%)	695

Table 6: Number of informative phrasal terms

Category	Phrasal terms
Informative	転送 レート 制御(transmission rate control),フロー 制御 (flow control), レート 制御 (rate control)
Neutral	マルチ キャスト 通信(multicast communication),マルチ キャスト (multicast),
Destructive	研究 動向(research trend),部分 的(partial),関連 性(relatedness),送信者 側 (sender side),複数 データ(multiple data),マルチ キャスト 環境 (multicast environment),マルチメディア データ(multimedia data),受信 者 (receiver)

Table 7 : Examples of phrasal terms in three categories from NACSIS topic 31

2.3. Three Categories of Phrasal Terms

The following three categories of phrasal terms in view of possible contribution to retrieval effectiveness are proposed from the previous discussion.

1) Informative phrasal terms : MI(occ,rel) > Σ MI(occ of constituent single words ,rel).

2) Neutral phrasal terms : Σ MI(occ of constituent single words ,rel) > MI(occ,rel) >= 0.

3) Destructive phrasal terms : MI(occ,rel) < 0.

For example, Table 7 shows phrasal terms extracted from all fields of topic 31 in NACSIS test collection 1,and classified accordingly.

2.4. Weight Ratio of Phrasal Terms

Retrieval status values are computed as a linear combination of each term weight, which is the product of the query weight and the document weight of the term. Using atn weighting in the SMART system for the same setting as the runs reported in Table 1, for each query term, the sums of weights of each query term are computed and for each query weight sum, ratio of informative phrasal terms and destructive phrasal terms are also computed. Macro-averaged ratios of informative phrasal terms and destructive phrasal terms are shown in Table 8. Still, short queries seem to contain better phrases in the ratio despite the fact that no consistent effectiveness for retrieval performance is observed.

2.5. Correlation between phrasal term weight ratio and performance difference

For each runs against the 53 test topic set both with short queries and long queries, correlation between query-by-query performance difference and query-by-query weight ratio of both informative and destructive phrasal term weight ratio are examined. Performance difference is measured by non-interpolated average precision and when the supplemental phrasal term run performs better a positive value is given as we have seen in Table 1.

Table 9 shows the Pearson's correlation coefficient between performance difference and each weight ratio as well as and difference between weight ratios.

	Average weight ratio of informative phrases	Number of topics Containing informative phrases	Average weight ratio of destructive phrases	Number of topics containing destructive phrases
Short query	8.59%	25	10.40%	26
Long query	6.47%	47	16.14%	53

Table 8 : Weight ratio of phrasal terms (macro-averaged for 53 topics)

	Informative phrasal term weight ratio(A)	Destructive phrasal term weight ratio(B)	(A)-(B)
Short query	0.12	-0.05	0.11
Long query	0.02	-0.05	0.04

Table 9 : Pearson's correlation between performance difference and phrasal term weight ratio

A positive correlation coefficient for informative phrasal terms and a negative correlation coefficient for destructive phrasal terms are observed as is expected, although the coefficient values are very small.

Given a topic set, a document collection and relevance judgements, we are able to know which terms are good (and possibly how good they are) for retrieval performance but to explain slight performance difference between different indexing strategies seems to be much more difficult.

Short queries contain relatively better phrasal terms even though absolute number of such terms is smaller than longer queries. But utilizing such phrasal terms does not always lead to performance improvement in macro-averaged precision-recall basis evaluation.

3. Topic Deviation

What we mean by topic deviation is a phenomenon that is similar to query drift caused by relevance feedback, but is incurred by some over-weighted supplemental phrasal terms. Terms representing some concepts in the topic are over-weighted consequently the search results are inclined to these concepts.

We verified short queries where supplemental phrasal terms caused considerable degradation (difference in average precision is more than 20%) and listed phrasal terms caused such degradation in Table 10.

As we can see, not only the neutral phrases in topics 50, 62 and 77, but also adding only informative phrases caused degradation as in topic 76.

<description> field of topic 76 is translated as follows:

"(I want to know about) methods for interference detection between polyhedral representations."

This topic consists of two concepts namely "interference detection" and "polyhedral representation" and the supplemented phrasal tem "多面体 間"(between polyhedral) is part of the second concept.

Retrieval effectiveness depends on a subtle balance of weighting on each concept, especially in short queries, and redundant terms or over-weighted terms cause the scoring function to loose such balances.

Conclusions

Effects of phrasal indexing in view of different length of queries are observed in the experiments using NACSIS test collection 1, the first large scale test collection for Japanese information retrieval.

Our observations and conclusions are as follows:

1) Distribution characteristics of phrasal terms as well as single word terms are examined plotting each term's MI(occ,rel) as function of log(O(occ)).

2) Distribution characteristics of phrasal terms are similar to single word terms but their frequencies are much smaller than single words.

3) Generally phrasal terms are comparably good discriminators of relevant documents, if not superior, as single words are.

4) In supplemental phrasal indexing, good discriminator terms are not always effective for retrieval performance but only some phrasal terms are informative and possibly effective.

5) Informative, neutral and destructive phrasal terms are defined by means of MI(occ,rel).

6) Correlation between performance difference and weight ratio of informative/destructive terms is examined and a very week correlation is observed.

Topic	Term	p(occ)	p(occ|rel)	p(occ|~rel)	log(p(occ|rel) /p(occ))	Category
34	改良 方法 (improvement method)	0.000129	0	0.000129	-6	Destructive
50	人工 知能 (artificial intelligence)	0.002346	0.388889	0.002305	5.11063	Neutral
60	教育 問題 (educational issues)	0.000006	0	0.000006	-6	Destructive
60	占領 期 (occupation period)	0.000012	0.222222	0.000006	9.848081	Informative
60	教育 事情 (educational situation)	0.000009	0	0.000009	-6	Destructive
62	生涯 学習 (life-long learning)	0.00044	0.285714	0.000435	6.475054	Neutral
76	多面体 間 (between polyhedral)	0.000023	0.076923	0.000021	8.094061	Informative
77	点字 翻訳 (braille transcription)	0.000029	0.166667	0.000026	8.644108	Neutral
78	哺乳 動物 (mammals)	0.000414	0	0.000414	-6	Destructive
78	不死 化 (immortalize)	0.000065	0.666667	0.000053	9.241945	Informative

Table 10: Phrasal terms in degraded topics by supplemental phrases

7) Explaining effectiveness of each query term is not sufficient for explaining effectiveness of phrasal indexing. Even good discriminator terms may hurt the retrieval effectiveness.

This research is by no means conclusive but a starting point of a longer project that hopefully leads to a new weighting scheme to replace current empirical down-weighting approach for supplemental phrasal terms.

Acknowledgements

The author thanks NACSIS R&D department for providing us of NACSIS test collection 1. We participated in the NTCIR workshop utilizing NACSIS test collection 1 (preliminary version) that is developed by NACSIS R&D department, thanks to understanding of academic societies (http://www.rd.nacsis.ac.jp/~ntcadm/thanks1-en.html) who provided the data.

References

[1] Fujita, S. (1999). Notes on Phrasal Indexing: JSCB Evaluation Experiments at NTCIR AD HOC, NTCIR Workshop 1, Tokyo, 101-108.

[2] Fujita, S. (2000). Evaluation of Japanese Phrasal Indexing with a Large Test Collection, RIAO2000 Conference proceedings, Paris, 1089-1098.

[3] Greiff, W.R. (1998). A Theory of Term Weighting Based on Exploratory Data Analysis, SIGIR '98, Melbourne, 11-19.

[4] NTCIR. (1999). http://www.rd.nacsis.ac.jp/~ntcadm/index-en.html .

[5] Pickens, J. Croft, W.B. (2000) An Exploratory Analysis of Phrases in Text Retrieval, RIAO2000 Conference proceedings, Paris, 1179-1195.