Reflections on "Aboutness"
TREC-9 Evaluation Experiments at Justsystem

Sumio FUJITA
JUSTSYSTEM Corporation
Brains park, Tokushima, 771-0189 JAPAN
+81-88-666-1000
Sumio_Fujita@justsystem.co.jp

ABSTRACT

TREC-9 evaluation experiments at the Justsystem site are described with a focus on "aboutness" based approach in text retrieval.

Experiments on the effects of supplemental noun phrase indexing, pseudo-relevance feedback and reference database feedback in view of the effect of various length of queries are reported.

The results show that pesudo-relevance feedback is always effective while reference database feedback is effective only with very short queries.

We reconfirmed that supplemental phrasal indexing is more effective with longer queries.

Keywords

Aboutness, Supplemental Phrasal indexing, phrasal terms, pseudo-relevance feedback, reference database, vector space model.

1. INTRODUCTION

Automatic indexing of modern information retrieval systems typically adopts bag-of-word representation, in which each word is considered as a dimension of the vector representing an information item, as internal representation of "aboutness". It is well known that such simple representation usually performs, as well as, if not better than, some more sophisticated ones according to empirical evaluations.

Grammatical relations or functional words are normally considered as neutral in view of thematic discrimination of text documents. On the other hand, content words(or lexemes, if we need to be more attentive for linguistic terminology) are semiologically meaningful units in language systems which refer to conceptual/substantial entities or relations in the subject domain described by the documents. It is plausible that the author of documents and the user submitting search requests share the same terminology when describing the subject concept in question either in their documents or in queries. The notion of "aboutness" is considered as a set of terms evoking a subject concept, which is hopefully shared by many people including authors, indexers and users of the system.

2. "ABOUTNESS"

The concept of "aboutness" plays an essential role in modern information retrieval technologies where "author's aboutness" [Ingwersen 93] is extracted automatically from text documents by automatic indexing procedures.

2.1 "Aboutness" as Representation of Information Objects

The basic hypothesis behind our TREC-9 strategies is that the "aboutness" of a subject topic consists of "foreground" part and "background" part and terms belong to either one of them. This distinction is inspired by the metaphor of "aboutness" of visual information items. People are clearly distinguishing foreground images from background ones when talking about "aboutness" of for example picture images. A foreground image might be a person or some objects located in the center of the picture and constitute the motif of the picture. Background images can help to identify the scene where the motif image is located and sometimes clue images are hidden in background when some implicit information is given in the picture.

In text retrieval, we can consider concepts that directly related to the motif as foreground and concepts that simply constitute the scene of the motif as background.

The term weighting should accordingly take this into consideration so that the terms that belong to "foreground aboutness" should be more weighted than "background aboutness".

Foreground terms are mainly extracted from <title> or <description> fields of topic description.

A stratified automatic feedback strategy is adopted in order to extract mainly terms of "background aboutness" both from the target document database(wt10g) and a reference database(TREC CD4&5).

2.2 Single words as a minimum unit of "aboutness"

Single words are indexed as basic units of "aboutness" but also noun phrases are extracted as supplemental indexing units.

For example, from the TREC topic 468 the following terms are extracted:

PH(incandescent light bulb)

PH(incandescent light),PH(light bulb)

incandescent, light, bulb

Longer phrases have normally more specific reference consequently they seem to focus more on foreground part of subject description while a set of constituent single word terms are referring to the subject as if it is on background.

Changing relative weighting of phrases against single word terms, "aboutness" of the query, especially its focusing strength can be calibrated without introducing any semantic hierarchy from thesauri.

We observed the correlation between query length and effectiveness gained by supplemental phrasal indexing [Fujita 00a, Fujita 00b]. It is still in open question that such a difference of phrasal term effectiveness in different length of queries can be explained from the difference of "aboutness".

2.3 Reference Database as a Substitute for a Thesaurus

Since web queries are typically short and do not contain enough terms to discriminate documents, query expansion is desirable for the better results in TREC style evaluations.

For an automatic query expansion purpose, typically synonymous words from a thesaurus are utilized.

In Japanese text retrieval experiments, we once tried such a strategy and observed consistent but small improvement with a newspaper article database [Fujita 99b].

Such an approach is problematic since preparing and maintaining thesauri is not an easy task either for an open domain or a closed domain.

Another problem of utilizing pre-coded thesauri for query expansion is that synonymous relations described in thesauri are not necessarily mean equivalence as a query term. Semantic equivalence relations in lexicon level do not necessarily mean equivalence in subject concepts of retrieved documents.

Instead of such a semantic approach, documents themselves, which represent author's "aboutness" can be utilized as the source of query expansions. The technique is similar to pseudo-relevance feedback procedures, that is frequently used in TREC experiments but the database in pilot search is not identical to the retrieval target database itself. Since many web documents are terminologically so poor that it is natural to refer to other text sources for term extraction.

A reference database can be either general domain databases like newspaper or a specific domain database depending on the retrieval task in question.

In the case of web retrieval, a newspaper database seems to be appropriate, since it is open domain retrieval and the reference databases preferably cover the any subjects that might be in test topics. Only newspapers and encyclopaedia seem to possess such a broad coverage of content documents.

2.4 Another source of "aboutness": Anchor Text of Hyperlinks

When we ask what a page is talking about, sometimes anchor texts (or link texts, the texts on which a hyperlink is set) indicate exact and very short answer.

The anchor text is typically an explanation or denotation of the page that is linked to. Some commercial based search engines are utilizing such information for advanced searches [Altavista]. We treat anchor texts literally as the part of the linked document.

In total, 6,077,878 anchor texts are added to 1,173,189 linked pages out of 1,692,096 pages in the wt10g data set. So 69% document pages in the data set are attributed anchor text information on top of original page information.

3. SYSTEM DESCRIPTION

For the TREC-9 Web track experiments, we utilized the engine of Justsystem ConceptBase Search™ version 2.0 as the base system.

A dual Pentium III™ server (670MHz) running Windows NT™ server 4.0 with 1024MB memory and 136GB hard disk is used for experiments.

The document collections are indexed wholly automatically, and converted to inverted index files of terms.

3.1 Term Extraction

Queries and documents in target databases are analyzed by the same module that decomposes an input text stream into a word stream and parses it using simple linguistic rules , in order to compose possible noun phrases.

Extracted units are single word nouns as well as simple linguistic noun phrases that consist of a sequence of nouns or nouns preceded by adjectives.

3.2 Vector Space Retrieval

Each document is represented as a vector of weighted terms by tf*idf in inverted index files and the query is converted in similar ways.

Similarity between vectors representing a query and documents are computed using the dot-product measure, and documents are ranked according to decreasing order of RSV.

OKAPI BM25 function is utilized as TF part of weighting function [Robertson 94, Robertson 95] so that the retrieval process can be considered as probabilistic ranking.

3.3 Passage Retrieval

Since some pages are extremely long in the wt2g data set, we became aware of using passages rather than whole pages as the indexing unit is appropriate for the sake of retrieval effectiveness.

Passage delimiting is done by the manner that each passage becomes similar length rather than finding paragraph boundary.

3.4 Phrasal Indexing and Weighting

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 terms are treated in the same way, both as independent terms, where the frequency of each term is counted independently based on its occurrences .

As we indicated in [Fujita 99a, Fujita 00a], phrasal terms are over-weighted with normal scoring function. We evaluated the following three methods:

1) Empirical down-weighting method [Fujita 99a]
2) Fagan's method [Fagan 87]
3) Approximation to Robertson's method [Robertson 97]

As it performed always better than other methods in the pre-submission experiments, we adopted down-weighting approach although it requires empirical parameter tuning.

Another advantage of down-weighting approach is that the query specificity can be calibrated changing down-weighting parameters when enough phrasal terms are provided in the query.

3.5 Pseudo-Relevance Feedback and Reference Database Feedback

Automatic feedback strategy using pseudo-relevant documents is adopted for automatic query expansion.

The system submits the first query generated automatically from topic descriptions against the target or reference database, and considers the top n documents from relevant ranking list as relevant.

The term selection module extracts salient terms from these pseudo-relevant documents and adds them to the query vector.

Then the expanded query vector is submitted against the target database again and the final relevance ranking is obtained.

The whole retrieval procedure is as follows:

1) Automatic initial query construction from the topic description
2) 1st pilot search submitted against a reference database
3) Term extraction from pseudo-relevant documents and feedback
4) 2nd pilot search submitted against the target database
5) Term extraction from pseudo-relevant documents and feedback
6) Final search to obtain the final results

3.6 Term Selection

Each term in example documents are scored by some term frequency and document frequency based heuristics measures described in [Evans 93].

The terms thus scored are sorted in decreasing order of each score and cut off at a threshold determined empirically.

In effect, the following parameters in feedback procedures should be decided:

1) How many documents to be used for feedback?

2) Where to cut off ranked terms?

3) How to weight these additional terms?

These parameters are carefully adjusted using TREC-8 queries (topic 401-450), wt2g data set and their relevance judgement provided by NIST and 4 parameter sets for official runs are decided.

Run tag	Query	Link	Ref	Avg. Prec	R-Prec
jscbt9wcs1	VS	No	Yes	0.2011	0.2175
jscbt9wls1	VS	Yes	Yes	0.2000	0.2219
jscbt9wls2	VS	Yes	No	0.1838	0.2027
jscbt9wcl1	Long	No	Yes	0.2687	0.2841
jscbt9wll1	Long	Yes	Yes	0.2659	0.2812
jscbt9wll2	Long	Yes	No	0.2801	0.3054

Table 1: Performance of official runs

3.7 Spell Variation

Because of some spelling errors in "title" field texts of topic description, the system sometimes returned no document or few in very short query runs. In such a case, the initial queries are expanded automatically by generated spell variations.

The procedure consists of looking for similar words in the word lists extracted from the database. Spelling similarity is measured by a combination of uni-gram, bi-gram and tri-gram matching scores.

4. EXPERIMENTS

We submitted six automatic runs as follows:

jscbt9wcs1: Content only, very short query run with parameter set s1

jscbt9wls1: Link, very short query run with parameter set s1

jscbt9wls2: Link, long query run with parameter set s2

jscbt9wcl1: Content only, long query run with parameter set l1

jscbt9wll1: Link, long query run with parameter set l1

jscbt9wll2: Link, long query run with parameter set l2

As for the link run evaluation, we adopted "anchor text" of hyperlink information as some web search sites do.

The experiments are designed to measure effects of phrasal term indexing, pseudo-relevance feedback and reference database feedback with regards to different query types.

From our experience in NTCIR-1 experiments for Japanese text retrieval, we are paying attention to the relation between the effectiveness of elementary techniques and the query length.

We observed that performance gain by the pseudo-relevance feedback tend to be large when the query is shorter in NTCIR-1 experiments. It is easily understood that longer queries contain already so good terms that the feedback could no more find better terms in addition.

It seems more difficult to explain why supplemental phrasal indexing is more effective with longer queries.

4.1 Very Short Query Experiments

Very short query run using only "title" fields of topic description is recommended for all the sites.

The following settings are examined:

1. Content only, single words + phrases
2. Link, single words + phrases
3. Content only, single words
4. Link, single words

For each setting, combination of with/without reference database feedback and with/without pseudo-relevance feedback are examined with the same parameter set: s1, for the convenience of comparison. Results of 16 runs in total are compared in Table 2.

Since initial queries are very short (in average, 2.1 single word terms and 0.7 phrasal terms, maximum 5 single word terms and 3 phrasal terms , minimum 0 single word terms and 0 phrasal terms) and they do not contain enough terms, the automatic feedback procedure contributes to 4.5% to 7.5 % of consistent improvements in average precision in all cases.

The final queries contain 44.1 single word terms and 31.0 phrasal terms in average (maximum 138 single word terms and 176 phrasal terms, minimum 0 single word terms and 0 phrasal terms).

The improvement gained by the combination of a pseudo-relevance feedback and reference database feedback is 15.8% for content only run and 17.0% for link run.

Run description	Ref	PFB	AvgPrec	R-Prec
Content only / very short / SW + phrases	Yes	Yes	0.2028	0.2185
Content only / very short / SW + phrases	Yes	No	0.1893	0.2267
Content only / very short / SW + phrases	No	Yes	0.1849	0.2135
Content only / very short / SW + phrases	No	No	0.1751	0.2020
Link / very short / SW + phrases	Yes	Yes	0.2018	0.2228
Link / very short / SW + phrases	Yes	No	0.1927	0.2228
Link / very short / SW + phrases	No	Yes	0.1854	0.2082
Link / very short / SW + phrases	No	No	0.1725	0.1919
Content only / very short / Single words only	Yes	Yes	0.1864	0.1949
Content only / very short / Single words only	Yes	No	0.1714	0.1987
Content only / very short / Single words only	No	Yes	0.1763	0.2022
Content only / very short / Single words only	No	No	0.1683	0.2025
Link / very short / Single words only	Yes	Yes	0.1863	0.1976
Link / very short / Single words only	Yes	No	0.1732	0.1922
Link / very short / Single words only	No	Yes	0.1726	0.1948
Link / very short / Single words only	No	No	0.1693	0.1983

Table 2: Performance comparison (Very Short Query, s1 parameter set)

Supplemental phrasal indexing runs perform better in average precision both with/without pseudo-relevance feedback and with/without reference database feedback.

But without any feedback, single word runs are better in R-precision.

Again we confirmed the situation observed in Japanese text retrieval workshop NTCIR-1 [Fujita 99a], i.e. effectiveness of phrasal indexing is not clear when the queries are short.

Effectiveness of link run is not clear as well.

4.2 Long Query Experiments

Long query experiments examined queries automatically constructed from all fields in topic description.

Since TREC topic descriptions have a stratified explanation of topics in the sense that the subject explanations are iterated in different styles. Shorter fields contain only terms of “foreground aboutenss” and longer fields contain terms of “background aboutness” as well as terms of “foreground aboutness”. It is important to adjust weighting for each term according to its “foregroundness” in the “request aboutness”.

We adjusted term weights according to the fields in which the term appeared since this might be a good measure for term “foregroundness”.

The same runs as very short query are examined:

1. Content only, single words + phrases
2. Link, single words + phrases
3. Content only, single words
4. Link , single words

The initial queries contain 11.6 single word terms and 3.46 phrasal terms in average (maximum 18 single word terms and 9 phrasal terms, minimum 5 single word terms and 0 phrasal terms) and the final queries contain 76.9 single word terms and 53.6 phrasal terms in average (maximum 239 single word terms and 218 phrasal terms, minimum 25 single word terms and 5 phrasal terms).

Table 3 shows the results. Supplemental phrasal runs are consistently better than single word term runs both in average precision and R-precision.

Since initial queries are longer and they contain terms of “background aboutness”, performance improvements given by automatic feedback are comparatively smaller (o.3%-6.5%) than in very short query experiments (4.5%-7.5%).

No search effectiveness improvement by introducing feedback from a reference database is observed.

We reconfirmed our observation from Japanese text retrieval experiments that the phrasal term indexing is effective only with enough long initial topic description containing a certain number of phrases as well as single words, otherwise its effect is rather incidental.

Run description	Ref	PFB	AvgPrec	R-Prec
Content only / Long / SW + phrases	Yes	Yes	0.2666	0.2784
Content only / Long / SW + phrases	Yes	No	0.2612	0.2940
Content only / Long / SW + phrases	No	Yes	0.2771	0.3067
Content only / Long / SW + phrases	No	No	0.2649	0.3043
Link / Long / SW + phrases	Yes	Yes	0.2650	0.2861
Link / Long / SW + phrases	Yes	No	0.2642	0.2962
Link / Long / SW + phrases	No	Yes	0.2801	0.3054
Link / Long / SW + phrases	No	No	0.2631	0.2942
Content only / Long / Single words only	Yes	Yes	0.2486	0.2518
Content only / Long / Single words only	Yes	No	0.2516	0.2793
Content only / Long / Single words only	No	Yes	0.2568	0.2883
Content only / Long / Single words only	No	No	0.2456	0.2762
Link / Long / Single words only	Yes	Yes	0.2480	0.2538
Link / Long / Single words only	Yes	No	0.2534	0.2772
Link / Long / Single words only	No	Yes	0.2614	0.2882
Link / Long / Single words only	No	No	0.2449	0.2729

Table 3: Performance comparison (Long query, l2 parameter set)

As in the very short query runs, it is not clear at all if link runs are better or not than content only runs. In the pre-submission experiments with the wt2g database and TREC-8 topics, small but consistent improvement was observed, but it is not the case with the TREC-9 main web test set. We did not yet find enough reason for this.

5. CONCLUSIONS

TREC-9 experiments at Justsystem group are described.

The following conclusions are drawn from these experiments:

1) Phrasal indexing seems to be more effective when the query is longer.

2) Pseudo-relevance feedback always contributes to the performance especially when initial queries are very short.

3) Feedback from a reference database was effective with very short queries but not with long queries.

4) No reliable performance improvement utilizing anchor texts was observed in wt10g experiments. Sometimes it was effective but not always.

On the other hand, we need more experiments as well as careful observation on the effect of phrasal indexing with short queries.

It is also interesting to compare the effects of reference database feedback with query expansion by WordNet style pre-coded thesauri.

For the future work, it is desirable to introduce the distinction of foreground/background of “aboutness” in question answering task where identification of focus of the topic description is crucial.

6. ACKNOWLEDGMENTS

Our thanks to Mr. Toshiya Ueda and Mr. Tatsuo Kato for their assistance.

REFERENCES

[1] Altavista: http://doc.altavista.com/adv_search/ast_ma_clickhere.html

[2] Evans, D.A. and Lefferts, R.G., Grefenstette, G., Handerson, S.K., Hersh, W.R., and Archbold, A.A., CLARIT TREC Design, Experiments and Results, in Proceedings of the First Text REtrieval Conference(TREC-1), NIST Special Publication 500-207, Washington D.C., 1993, 494-501.

[3] Fagan, J.L. Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-syntactic Methods, Ph.D Thesis,

Dept. of Computer Science, Cornell University, Sept. 1987.

[4] Fujita, S. Notes on Phrasal Indexing—JSCB Evaluation Experiments at NTCIR AD HOC, in Proceedings of NTCIR-1 workshop, 1999.

[5] Fujita, S. Notes on Phrasal Indexing: JSCB Evaluation Experiments at IREX-IR, in Proceedings of IREX Workshop, Tokyo, 1999, 45-51.

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

[7] Fujita, S. Discriminative Power and Retrieval Effectiveness of Phrasal Indexing Terms, in Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval, Hong Kong, 2000, 47-55.

[8] Ingwersen, P. Information Retrieval Interaction, Taylor Graham Publishing, London, 1993.

[9] Lewis, D. Representation and Learning in Information Retrieval, Ph.D Thesis, Dept. of Computer and Information Science, University of Massachusetts, Feb. 1992.

[10] Lewis, D. An Evaluation of Phrasal and Clustered representation on a Text Categorization Task, in Proceedings of the Fifteenth Annual International ACM SIGIR Conference(Copenhagen, June 1992), ACM Press, 37-50.

[11] Robertson, S.E. and Walker S. Some Simple Effective Approximations to the 2Poisson Model for Probabilistic Weighted Retrieval, in Proceedings of the Seventeenth Annual International ACM SIGIR Conference(Dublin, July 1994), Springer-Verlag, 232-241.

[12] Robertson, S.E., Walker S., Jones S., Hancock-Beaulieu, M.M., Gatford, M. Okapi at TREC-3, in Proceedings of the Third Text REtrieval Conference(TREC-3), NIST Special Publication 500-225, Washington D.C., 1995, 109-126.

[13] Robertson, S.E. and Walker S. On relevance weights with little relevance information, in Proceedings of the 20th Annual International ACM SIGIR Conference(Philadelphia, July 1997), ACM Press, 16-24.

Appendix A.

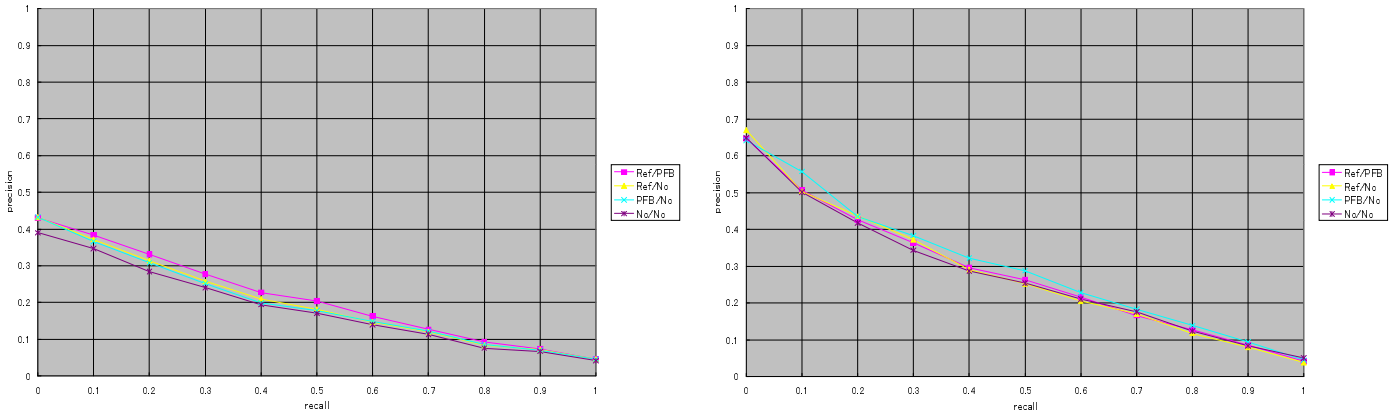

Figure 1: Recall-precision curves of supplemental phrasal runs
Left: Content only very short runs, Right: Link long runs

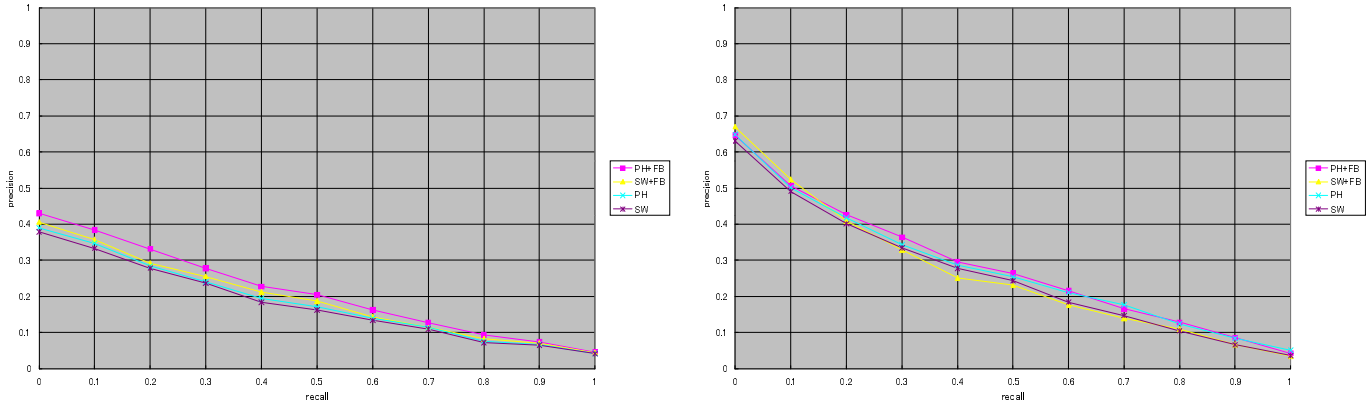

Figure 2: Recall-precision curves of supplemental phrasal runs vs single word runs with/without feedback Left: Content only very short runs Right: Link long runs