Term Distillation in Patent Retrieval

Hideo Itoh Hiroko Mano Yasushi Ogawa
Software R&D Group, RICOH Co., Ltd.
1-1-17 Koishikawa, Bunkyo-ku, Tokyo 112-0002, JAPAN
{hideo,mano,yogawa}@src.ricoh.co.jp

Abstract

In cross-database retrieval, the domain of queries differs from that of the retrieval target in the distribution of term occurrences. This causes incorrect term weighting in the retrieval system which assigns to each term a retrieval weight based on the distribution of term occurrences. To resolve the problem, we propose "term distillation", a framework for query term selection in cross-database retrieval. The experiments using the NTCIR-3 patent retrieval test collection demonstrate that term distillation is effective for cross-database retrieval.

1 Introduction

For the mandatory runs of NTCIR-3 patent retrieval task [1], participants are required to construct a search query from a news article and retrieve patents which might be relevant to the query. This is a kind of cross-database retrieval in that the domain of queries (news article) differs from that of the retrieval target (patent) (Iwayama et al., 2001).

Because in the distribution of term occurrences the query domain differs from the target domain, some query terms are given very large weights (importance) by the retrieval system even if the terms are not appropriate for retrieval. For example, the query term "president" in a news article might not be effective for patent retrieval. However, the retrieval system gives the term a large weight, because the document frequency of the term in the patent genre is very low. We think these problematic terms are so many that the terms cannot be eliminated using a stop word dictionary.

In order to resolve the problem mentioned above, we propose "term distillation" which is a framework for query term selection in cross-database retrieval. The experiments using the NTCIR patent retrieval test collection demonstrate that term distillation is effective for cross-database retrieval.

2 System description

Before describing our approach, we give a short description on our retrieval system. For the NTCIR-3 experiments, we revised query processing although the framework is the same as that of NTCIR-2 (Ogawa and Mano, 2001). The basic features of the system are as follows :

- Effective document ranking with pseudo-relevance feedback based on Okapi's approach (Robertson and Walker, 1997) with some improvements.
- Scalable and efficient indexing and search based on the inverted file system (Ogawa and Matsuda, 1999)
- Originally developed Japanese morphological analyzer and normalizer for document

[1] http://research.nii.ac.jp/ntcir/ntcir-ws3/patent/

indexing and query processing.

The inverted file was constructed for the retrieval target collection which contains full texts of two years' Japanese patents. We adopted character n-gram indexing because it might be difficult for Japanese morphological analyzer to correctly recognize technical terms which are crucial for patent retrieval.

In what follows, we describe the full automatic process of document retrieval in the NTCIR-3 patent retrieval task.

1. Query term extraction
   Input query string is transformed into a sequence of words using the Japanese morphological analyzer. Query terms are extracted by matching patterns against the sequence. We can easily specify term extraction using some patterns which are described in regular expression on each word form or tag assigned by the analyzer. Stop words are eliminated using a stop word dictionary. For initial retrieval, both "single terms" and "phrasal terms" are used. A phrasal term consists of two adjacent words in the query string.

2. Initial retrieval
   Each query term is submitted one by one to the ranking search module, which assigns a weight to the term and scores documents including it. Retrieved documents are merged and sorted on the score in the descending order.

3. Seed document selection
   As a result of the initial retrieval, top ranked documents are assumed to be pseudo-relevant to the query and selected as a "seed" of query expansion. The maximum number of seed documents is ten.

4. Query expansion
   Candidates of expansion terms are extracted from the seed documents by pattern matching as in the query term extraction mentioned above.

   Phrasal terms are not used for query expansion because phrasal terms may be less effective to improve recall and risky in case of pseudo-relevance feedback.

   The weight of initial query term is re-calculated with the Robertson/Spark-Jones formula (Robertson and Sparck-Jones, 1976) if the term is found in the candidate pool.

   The candidates are ranked on the Robertson's Selection Value (Robertson, 1990) and top-ranked terms are selected as expansion terms.

5. Final retrieval
   Each query and expansion term is submitted one by one to the ranking search module as in the initial retrieval.

3 Term distillation

In cross-database retrieval, the domain of queries (news article) differs from that of the retrieval target (patent) in the distribution of term occurrences. This causes incorrect term weighting in the retrieval system which assigns to each term a retrieval weight based on the distribution of term occurrences. Moreover, the terms which might be given an incorrect weight are too many to be collected in a stop word dictionary.

For these reasons, we find it necessary to have a query term selection stage specially designed for cross-database retrieval. We define "term distillation" as a general framework for the query term selection.

More specifically, the term distillation consists of the following steps :

1. Extraction of query term candidates
   Candidates of query terms are extracted from the query string (news articles) and pooled.

2. Assignment of TDV (Term Distillation Value)
   Each candidate in the pool is given a TDV which represents "goodness" of the term to retrieve documents in the target domain.

3. Selection of query terms
The candidates are ranked on the TDV and top-ranked n terms are selected as query terms, where n is an unknown constant and treated as a tuning parameter for full-automatic retrieval.

The term distillation seems appropriate to avoid falling foul of the "curse of dimensionality" (Robertson, 1990) in case that a given query is very lengthy.

In what follows in this section, we explain a generic model to define the TDV. Thereafter some instances of the model which embody the term distillation are introduced.

3.1 Generic Model

In order to define the TDV, we give a generic model with the following formula.

$$TDV = QV \cdot TV$$

where QV and TV represent the importance of the term in the query and the target domain respectively. QV seems to be commonly used for query term extraction in ordinary retrieval systems, however, TV is newly introduced for cross-database retrieval. A combination of QV and TV embodies a term distillation method. We instance them separately as bellow.

3.2 Instances of TV

We give some instances of TV using two probabilities p and q, where p is a probability that the term occurs in the target domain and q is a probability that the term occurs in the query domain. Because the estimation method of p and q is independent on the instances of TV, it is explained later. We show each instance of TV with the id-tag as follows:

TV0 : Zero model
$TV = constant = 1$

TV1 : Swet model (Robertson, 1990)
$TV = p - q$

TV2 : Naive Bayes model
$TV = \frac{p}{q}$

TV3 : Bayesian classification model
$TV = \frac{\alpha \cdot p}{\alpha \cdot p + (1-\alpha-\epsilon) \cdot q + \epsilon}$
where α and ϵ are unknown constants.

TV4 : Binary independence model (Robertson and Sparck-Jones, 1976)
$TV = \log \frac{p(1-q)}{q(1-p)}$

TV5 : Target domain model
$TV = p$

TV6 : Query domain model
$TV = 1 - q$

TV7 : Binary model
$TV = 1 \quad (p > 0) \quad or \quad 0 \quad (p = 0)$

TV8 : Joint probability model
$TV = p \cdot (1 - q)$

TV9 : Decision theoretic model (Robertson and Sparck-Jones, 1976)
$TV = \log(p) - \log(q)$

3.3 Instances of QV

We show each instance of QV with the id-tag as follows:

QV0 : Zero model
$QV = constant = 1$

QV1 : Approximated 2-poisson model (Robertson and Walker, 1994)
$QV = \frac{tf}{tf+\beta}$
where tf is the within-query term frequency and β is an unknown constant.

QV2 : Term frequency model
$QV = tf$

QV3 : Term weight model
$QV = weight$
where $weight$ is the retrieval weight given by the retrieval system.

QV4 : Combination of QV1 and QV3
$QV = \frac{tf}{tf+\beta} \cdot weight$

QV5 : Combination of QV2 and QV3
$QV = tf \cdot weight$

4 Experiments on term distillation

Using the NTCIR-3 patent retrieval test collection, we conducted experiments to evaluate the effect of term distillation.

For query construction, we used only news article fields in the 31 topics for the formal run. The number of query terms selected by term distillation was just eight in each topic. As described in the section 2, retrieval was full-automatically executed with pseudo-relevance feedback.

The evaluation results for some combinations of QV and TV are summarized in Table 1, where the documents judged to be "A" were taken as relevant ones. The combinations were selected on the results in our preliminary experiments.

Each of "t", "i", "a" and "w" in the columns "p" or "q" represents a certain method for estimation of the probability p or q as follows :

t : estimate p by the probability that the term occurs in *titles* of patents. More specifically $p = \frac{n_t}{N_p}$, where n_t is the number of patent titles including the term and N_p is the number of patents in the NTCIR-3 collection.

i : estimate q by the probability that the term occurs in *news articles*. More specifically $q = \frac{n_i}{N_i}$, where n_i is the number of articles including the term and N_i is the number of news articles in the IREX collection ('98-'99 MAINICHI news article).

a : estimate p by the probability that the term occurs in *abstracts* of patents. More specifically $p = \frac{n_a}{N_p}$, where n_a is the number of patent abstracts in which the term occurs.

w : estimate q by the probability that the term occurs in the *whole* patent. More specifically $q = \frac{n_w}{N_p}$, where n_w is the number of patents in which the term occurs. We tried to approximate the difference in term statistics between patents and news articles using the conbination of "a" and "w" in the term distillation.

In Table 1, the combination of QV2 and TV0 corresponds to query term extraction without term distillation. Comparing with the combination, retrieval performances are improved using instances of TV except for TV7. This means the term distillation produces a positive effect.

QV	TV	p	q	AveP	P@10
QV2	TV4	t	i	0.1953	0.2645
QV2	TV9	t	i	0.1948	0.2677
QV5	TV3	t	i	0.1844	0.2355
QV2	TV3	t	i	0.1843	0.2645
QV0	TV3	t	i	0.1816	0.2452
QV2	TV6	t	i	0.1730	0.2258
QV2	TV2	t	i	0.1701	0.2194
QV2	TV3	a	w	0.1694	0.2355
QV2	TV0	-	-	0.1645	0.2226
QV2	TV7	t	i	0.1597	0.2065

Table 1: Results using article field

The best performance in the table is produced by the combination of QV2 (raw term frequency) and TV4 (BIM).

While the combination of "a" and "w" for estimation of probabilities p and q has the virtue in that the estimation requires only target document collection, the performance is poor in comparison with the combination of "t" and "i".

Although the instances of QV can be compared each other by focusing on TV3, it is unclear whether QV5 is superior to QV2. We think it is necessary to proceed to the evaluation including the other combinations of TV and QV.

5 Results in NTCIR-3 patent task

We submitted four mandatory runs. The evaluation results of our submitted runs are summarized in Table 2, where the documents judged to be "A" were taken as relevant ones.

These runs were automatically produced using both article and supplement fields, where each supplement field includes a short description on the content of the news article. Term distillation using TV3 (Bayes classification model) and query expansion by pseudo-relevance feedback were applied to all runs.

The retrieval performances are remarkable among all submitted runs. However, the effect

QV	*TV*	p	q	AveP	P@10
QV2	TV3	t	i	0.2794	0.3903
QV0	TV3	t	i	0.2701	0.3484
QV2	TV3	a	w	0.2688	0.3645
QV5	TV3	t	i	0.2637	0.3613

Table 2: Results in the NTCIR-3 patent task

of term distillation is somewhat unclear, comparing with the run with only supplement fields in Table 3 (the average precision is 0.2712). We think supplement fields supply enough terms so that it is difficult to evaluate the performance of cross-database retrieval in the mandatory runs.

6 Results on ad-hoc patent retrieval

In Table 3, we show evaluation results corresponding to various combinations of topic fields in use. The documents judged to be "A" were taken as relevant ones.

fields	AveP	P@10	Rret
t,d,c	0.3262	0.4323	1197
t,d,c,n	0.3056	0.4258	1182
d	0.3039	0.4032	1133
t,d	0.2801	0.3581	1100
t,d,n	0.2753	0.4000	1140
d,n	0.2750	0.4323	1145
s	0.2712	0.3806	991
t	0.1283	0.1968	893

Table 3: Results on ad-hoc patent retrieval

In the table, the fields "t", "d", "c", "n" or "s" correspond to title, description, concept, narrative or supplement respectively. As a result, the combination of "t,d,c" produced the best retrieval performance for a set of the formal run topics. Pseudo-relevance feedback had a positive effect except for the case using a title field only.

7 Conclusions

We proposed term distillation for cross-database retrieval. Using NTCIR-3 test collection, we evaluated this technique in patent retrieval and found a positive effect. We think cross-database retrieval can be applied to various settings including personalized retrieval, similar document retrieval and so on.

For the future work, we hope to apply term distillation to cope with vocabulary gap problems in these new settings. In addition, we think term distillation can be used to present query terms to users in reasonable order in interactive retrieval systems.

References

M. Iwayama, A. Fujii, A. Takano, and N. Kando. 2001. Patent retrieval challenge in NTCIR-3. *IPSJ SIG Notes*, 2001-FI-63:49–56.

Y. Ogawa and H. Mano. 2001. RICOH at NTCIR-2. *Proc. of NTCIR Workshop 2 Meeting*, pages 121–123.

Y. Ogawa and T. Matsuda. 1999. An efficient document retrieval method using n-gram indexing. *Trans. of IEICE*, J82-D-I(1):121–129.

S. E. Robertson and K. Sparck-Jones. 1976. Relevance weighting of search terms. *Journal of ASIS*, 27:129–146.

S. E Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. *Proc. of 17th ACM SIGIR Conf.*, pages 232–241.

S. E. Robertson and S. Walker. 1997. On relevance weights with little relevance information. *Proc. of 20th ACM SIGIR Conf.*, pages 16–24.

S. E. Robertson. 1990. On term selection for query expansion. *Journal of Documentation*, 46(4):359–364.