OKSAT at NTCIR-11 RecipeSearch
- Categorization and Expansion of Search Terms in Topics -

Takashi SATO
Information Processing Center
Osaka Kyoiku University
Kashiwara Osaka Japan
+81-72-978-3823
sato@cc.osaka-kyoiku.ac.jp

Shingo AOKI
Graduate School of Education
Osaka Kyoiku University
Kashiwara Osaka Japan
+81-72-978-3823
aoki@ss.osaka-kyoiku.ac.jp

Yuta MORISHITA
Department of Arts and Sciences
Osaka Kyoiku University
Kashiwara Osaka Japan
+81-72-978-3823
morishita@ss.osaka-kyoiku.ac.jp

ABSTRACT

Our group OKSAT submitted five runs for English and Japanese ad hoc recipe search (EN1 and JA1) subtasks of NTCIR-11 Cooking Recipe Search (RecipeSearch). For EN1, we tried to categorize search terms of topics. We also tried to expand search term for some runs we submitted. Analyzing experimental results, we observe the effectiveness of our method.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information filtering, Query formulation, Retrieval models, Search process, Selection Process.*

General Terms

Experimentation, Performance, Measurement.

Team Name

OKSAT

Subtasks

English ad hoc recipe search (EN1)
Japanese ad hoc recipe search (JA1)

Keywords

Information Retrieval, Cooking Recipe Search, Categorization of Search Terms, Expansion of Search Terms, Gram Base Index.

1. INTRODUCTION

Our group submitted five runs for English and Japanese ad hoc recipe search (EN1 and JA1) subtask of NTCIR-11 [1] Cooking Recipe Search (RecipeSearch) [2]. For runs of EN1, we tried to categorize search terms of topics. We also tried to expand search term for some runs. We do not expand search terms of JA1 topics because relatively detail information is obtained from the topic. Analyzing experimental results, we observe the effectiveness of our method.

2. OUTLINE OF OUR APPROACH

We searched corpus by the following procedure for English ad hoc recipe search (EN1) and Japanese ad hoc recipe search (JA1), and then we made runs.

(1) Extract fields from corpus and made four (EN1) or three (JA1) indices.
(2) Prepare search terms from topics to search indices of (1).
(3) Score search results of each index (2) using probabilistic model [3].
(4) Merge each scored results into a run.

Figure 1 shows the procedure above.

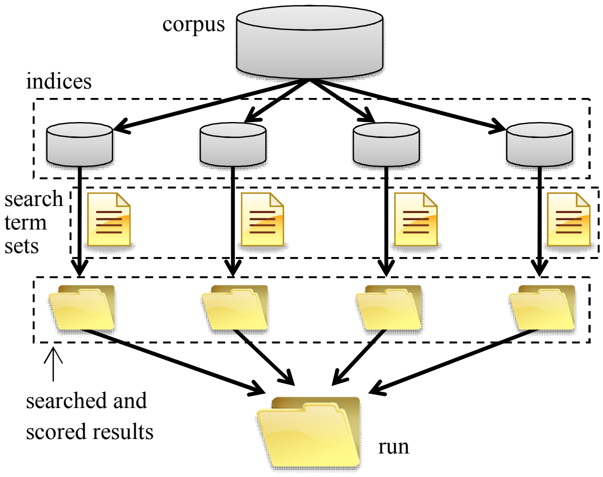

Figure 1. Procedure flow of our approach

The procedures of EN1 and JA1 were different each other in detail because fields of corpus and topics given from task organizer were different each other.

3. EN1

3.1 Indexing

From title, ingredientLines, preparationSteps and attributes field of English recipe corpus, we made **title**, **ingre**, **prep** and **attr** index correspondingly. We did not use the totalTimeInSeconds field of corpus. These were gram based indices [4][5][6], so arbitrary strings search was possible using them.

Table 1 shows specifications of computer we used. And Table 2 shows statistics of our EN indices and their creation time.

Table 1. Specifications of computer

CPU	Intel Core i5-4430@3.0GHz 4C/4T
MEM	8GB, DDR3-1600
O S	FreeBSD 8.4, 64bit
HDD	1TB, SATA 6GB/s, 64MB Cache

Table 2. Statistics of EN indices

	title	ingre	prep	attr
data size (MB)	2.77	30.3	64.4	3.19
index size (MB)	9.31	62.8	146	6.91
time (sec.)	1.12	11.5	25.9	.807

3.2 Categorization of Search Terms of Topic

We made search terms from a topic by the following procedures.

(1) Extract words from a topic.
(2) Categorize terms into four categories referring our recipe term database.

The categories are **ttl**, **ing**, **prp** and **att** intended to search title, ingre, prep and attr index of **2.1** respectively. Figure 2 ((1)) shows an example.

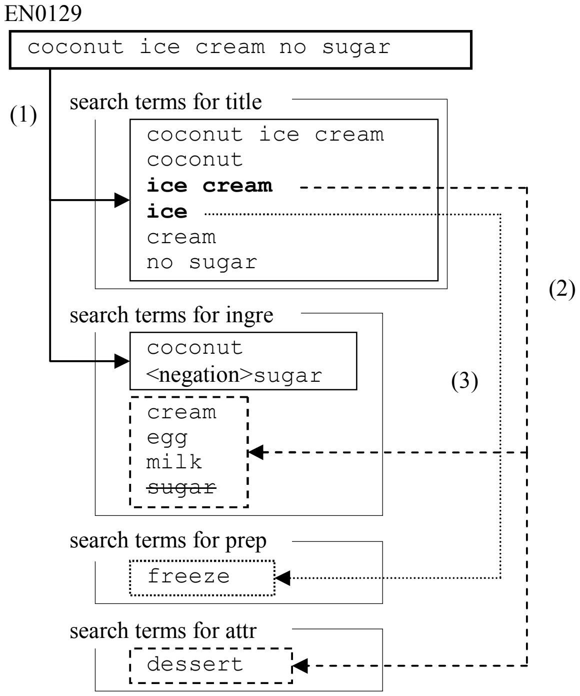

Figure 2. Categorization and expansion of search terms

3.3 Expansion of search Terms for index

We expanded search terms of **3.2** using words from example answer recipes and/or from the Internet search (Google, Wikipedia, Weblio, etc.). See Figure 3.

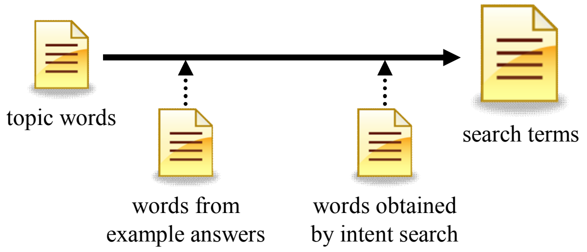

Figure 3. Expansion of search terms

Figure 2 ((2), (3)) shows an example again. Table 3 shows a part of our word expansion list. The expansion list was created manually about half of topics, and we tried to use this list to other topics by our expansion program.

Table 3. Part of word expansion list

type	source	expanded words
by grammar	strawberry	strawberries
ttl -> ing	bread	flour, baking powder
ing -> ing	fruit	apple, lemon, …
ttl -> att	cake, …	dessert

3.4 Searching, Scoring and Merging

We search four indices (title, ingre, prep, attr) of **3.1** by four search term sets (ttl, ing, prp, att) of **3.2** and **3.3**. Scoring each of search results using probabilistic model, we got four ranked document list namely title-ttl, ingre-ing, prep-prp and attr-att. We multiplied the ranked results by weight of 0.4, 0.4, 0.1, 0.1 in the order, and then we merged them into one list for a run.

3.5 Strength of Search Terms

Our system ranks document by probabilistic model as described in **3.4**. In order to enable Boolean type search, our system has the means of document filtering by the term strength defined below.

(1) Essential: should have the term
(2) Negation: should not have the term
(3) Essential + Parallel: at least one of grouped terms should appear in a document
(4) Negation + Not Negation: same as Negation if Not Negation terms appear in a document.

The negation search terms in ingre and prep are the topic words which are preceded or followed by words 'no', 'without', 'less' or 'free'. The essential search terms in title are the topic words which match terms of category 'title' in our recipe term database. And the essential search terms in ingre are the topic words which match terms of category 'ingre' in our recipe term database. The parallel search terms are the expanded words by using the expansion list of **3.3** from the essential search terms of a topic.

3.6 Submitted Runs

We added words from example answer recipes and/or from the Internet search as described in **3.3** to words from topic categorized as described in **3.2**. We made the following four runs by combinations of these search term sets.

OKSAT-EN1-TEST-01: words from topic only
OKSAT-EN1-TEST-02: topic + example answer
OKSAT-EN1-TEST-03: topic + internet search
OKSAT-EN1-TEST-04: topic + example answer +internet search

Table 4 shows time (searching, scoring and merging) and MAP (mean average precision)s of our submitted runs. These MAPs are obtained using NTCIREVAL [7] and they are the same as the official results for EN1 in [1].

Resolutions of search time were minutes because time was taken from time stamps of file accessed.

Table 4. Time and MAP of submitted EN1 runs

	time(min.)	MAP
OKSAT-EN1-TEST-01	5	0.6790
OKSAT-EN1-TEST-02	8	0.6999
OKSAT-EN1-TEST-03	9	0.7287
OKSAT-EN1-TEST-04	12	0.7499

3.7 Statistics of Topic Words

While processing topics, we observed some characteristics of topic words.

(1) Most words relate to title (dish name) and ingredientLines.
(2) There are words relate to cooking method (bake, fry, ...), cooking tool (casserole, slow cooker, ...), and manufacturing company.
(3) There are words relate to attribute such as season, region (country), time of the day, etc.
(4) Few words relate to cooking show, well-known cook, etc.
(5) 159 topics out of 500 topics in all have negation expression (... free, ... less, no ..., without ...). Most of them relate to ingrediantLines, however, a expression such as 'no bake' relates to preparetionSteps.

3.8 Topic by Topic Analysis

We show some easy and difficult topics for us.

(1) Topics in which titles (dishes) and/or ingredients, and/or cook tools are listed are easy. For example the following topics are such type.

EN0308: crock pot chicken mushrooms potatoes
EN0318: fish sticks without eggs
EN0322: baked potato with bacon and cheddar

We search titles (ingredients, cook tools) by title (ingre, prep) index with strength Essential or Negation of **3.5**.

(2) Topics in which include low fat, low calorie, etc. are difficult because we don't know these criterion.

EN0074: acorn squash low calorie soup
EN0118: diabetic low fat low cholesterol
EN0218: soba noodle salad low fat

(3) Topics which have few clues are difficult also.

EN0275: asian
EN0350: overnight breakfast

In those cases, we tried to search attr index.

4. JA1

4.1 Indexing

From recipe title and dish name fields in recipe_all file, we made **title** index. From material name field of recipe_material file, we made **mat** index. Finally from tag 1, tag 2, tag 3 and tag 4 fields in recipe_all file again, we made **tag** index. These JA indices were gram based ones also. Table 5 shows statistics of our JA indices and their creation time.

Table 5. Statistics of JA Indices

	title	mat	tag
data size(MB)	19.4	28.4	8.93
index size(MB)	31.9	44.4	12.6
time(sec.)	3.39	5.64	1.57

4.2 Relations Between Topic Field and Index

We made the following three search term sets from JA1 topic file.

(1) **ttl** from dishName and negation field
(2) **mt** from foodName
(3) **tg** from negation

The negation field in the topic was used twice. Because topic of JA1 consisted of plural fields unlike a case of EN1, we searched indices of **4.1** by corresponding search term sets above. We did not expand search terms in JA1 because relatively detail information was obtained from JA1 topic.

4.3 Searching, Scoring and Merging

We search three indices (title, mat, tag) of **4.1** by three search term sets (ttl, mt, tg) of **4.2**. Scoring each of search results using probabilistic model, we got three ranked document list namely title-ttl, mat-mt and tag-tg. We multiplied the ranked results by weight of 0.4, 0.4, 0.2 in the order, and then we merged them into one list for a run.

4.4 Submitted Run

As JA1 has no expanded search term sets we prepared different from EN1, we submitted only one run, namely OKSAT-JA1-TEST-01. Table 6 shows time and MAP of the run. This MAP is obtained using NTCIREVAL [7] and it is the same as the official results for JA1 in [1].

Table 6. Time and MAP of submitted JA1 run

	time(min.)	MAP
OKSAT-JA1-TEST-01	19	0.6849

5. EN1 vs. JA1

It is difficult to understand questioner's intension because topics of JA1 have plural fields. For example, foods listed in food name field in topics should be included or same in recipes. More ad hoc query similar to EN1 may help to compare language by language difference.

As extensions of JA1 topic, topics which intended to refer 'Standard Tables of Food Composition' might be interesting.

6. CONCLUSIONS

Our group submitted five runs for English and Japanese ad hoc recipe search (EN1 and JA1) subtask of NTCIR-11 Cooking Recipe Search (RecipeSearch). For EN1, while processing of topics, we made a categorization database from topic word and an expansion list for search terms. The expansion list was created manually about half of topics, and we tried to use this list to other topics by our expansion program. And then we tried to automate categorization and expansion of search terms using them.

7. ACKNOWLEDGMENTS

Our thanks to Yummly for allowing us to use 'Yummly Recipe Data v1', and to Rakuten Inc. for allowing us to use 'The Rakuten Data'.

8. REFERENCES

[1] H. Joho and K. Kishida, Overview of NTCIR-11 , in *Proceedings of the NTCIR-11* Conference, Tokyo, Japan, 2014.

[2] M. Yasukawa, F. Diaz, G. Druck, and N. Tsukada, Overview of NTCIR-11 Cooking Recipe Search Task, in *Proceedings of the NTCIR-11* Conference, Tokyo, Japan, 2014.

[3] S.E. Robertson and S. Walker, Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval, in *Proceedings of the 17th International Conference Research and Development in Information Retrieval*, pp. 232-241, 1994.

[4] T. Sato, Fast full text retrieval using gram based tree structure, in *Proceedings of the ICCPOL '97*, Vol.2, pp.572-577, 1997.

[5] T. Sato and K. Han, NTCIR-3 CLIR Experiments at Osaka Kyoiku University - Compression of Gram-based Indices -, in *Proceedings of the NTCIR-3*, Tokyo, December 2002.

[6] T. Sato, T. Satomoto, and K. Han, NTCIR-3 PAT Experiments at Osaka Kyoiku University -Long Gram-based Index and Essential Words -, in *Proceedings of the NTCIR-3*, Tokyo, December 2002.

[7] NTCIREVAL, http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html.