Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task

Makoto Iwayama[*], Atsushi Fujii[†], Noriko Kando[‡]
[*] Hitachi, Ltd., 1-280 Higashi-koigakubo, Kokubunji, Tokyo 185-8601, Japan
iwayama@crl.hitachi.co.jp
/ Tokyo Institute of Technology
[†] University of Tsukuba, 1-2 Kasuga, Tsukuba 305-8550, Japan
fujii@slis.tsukuba.ac.jp
[‡] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo 101-8430, Japan
kando@nii.ac.jp

Abstract

This paper describes Classification Subtask at NTCIR-5 Patent Retrieval Task. We perform two subtasks for patent classification using a multi-dimensional classification structure called "F-term (File Forming Term) classification system". The first one is Theme Categorization Subtask, where each participant classifies a patent into technological fields called themes. The second one is F-term Categorization Subtask, where each participant classifies a patent, whose theme has been given, into multifaceted categories called F-terms. We overview the designs of these subtasks, the test collections produced, and the evaluation results.

Keywords: *Patent Classification, F-term, theme, Patent Map*

1. Introduction

Companies that are trying to utilize their patents have to investigate the coverage of the patents in the targeting domain and clarify the advantages and disadvantages of the patents compared with competitors' patents. Patent maps help this kind of analysis by providing statistical information of patents from various perspectives.

In NTCIR-4 Patent Retrieval Task [2], we started a subtask (named as Feasibility Study Subtask) for creating a patent map which offers a bird's eye view of patents in a specific technological field. The patent map we targeted at was a two-dimensional matrix which summarizes given patents from two viewpoints, such as "problems to be solved" and "solutions". Figure 1 is an example of our targeting patent maps. In the map, columns ("crystalline", "reliability", "long life", etc.) are possible problems to be solved in the patents and lines ("structure of active layer", "electrode composition", etc) are possible solutions claimed in the patents. Patents in each cell solve the corresponding problem by the corresponding solution. For example, the patent "1998-107318" solves a problem about "reliability" of blue light-emitted diodes by an idea about "electrode composition".

problems to be solved

solutions		crystalline	reliability	long life	emission stability	emission intensity
	structure of active layer			1998-145000 1998-233554		
	electrode composition		1998-107318		1998-190063 1998-209498	1998-209495
	electrode arrangement		1998-215034 1998-223930	1998-242518	1998-173230 1998-209499 1998-256602	1998-242515 1998-270757
	structure of light emitting element	1998-135516 1998-242586 1998-247761		1998-135514 1998-256668		1998-012923 1998-247745 1998-256597

Figure 1. A patent map of blue light-emitted diodes.

Although this subtask revealed a couple of promising approaches for automatic patent map generation, there remained several problems especially about the evaluation issue. We had only six topics evaluated subjectively. In addition, the dimensions of patent maps were fixed to "problems to be solved" and "solutions" in all the topics. We have to note that the best combination of dimensions is different from a technological field to another.

In Classification Subtask at NTCIR-5 Patent Retrieval Task, we focused on the evaluation of patent classification by using a multi-dimensional classification structure called "F-term (File Forming Term) classification system"[1][3], which is used in the Japan Patent Office. F-term classification system has over 2,500 "themes" covering all the technological fields of patents. Patents under each theme can be classified from several viewpoints, such as purpose, function, effect, and so on. The collection of possible viewpoints varies from theme to theme. Each viewpoint defines a set of its possible elements and a pair of a viewpoint and its element is called "F-term". F-term classification system serves as an effective tool for narrowing down relevant patents in searching. It also helps for creating a two-dimensional patent

[1] http://www.ipdl.ncipi.go.jp/HELP/pmgs_en/database/format_summary.html#fterm

5B001		Detection and correction of errors					
AA	AA00	AA01	AA02	AA03	AA04	AA05	…
	CODES	. Parity	.. Multiple parity	. Error-correction codes (ECC)	.. Cyclic-redundancy check (CRC)	.. Single-bit error correction and double-bit error detection (SECDEC)	…
AB	AB00	AB01	AB02	AB03	AB04	AB05	…
	PURPOSE	. Error detection	. Error correction	. Code generation	.. Prediciton	. Decoding	…
AC	AC00	AC01	AC02	AC03	AC04	AC05	…
	MEANS	. Code operations	.. Tables	.. Counting	. Comparison	. Interleaving	…
AD	AD00	AD01	AD02	AD03	AD04	AD05	…
	ERROR LOCATION	. Arithmetic circuits	.. Decoders	. Memories	.. Magnetic tapes	. Interfaces	…
AE	AE00	AE01	AE02	AE03	AE04	AE05	…
	TYPES OF ERRORS	.. Program instructions	. Data	.. Multiple errors	...Burst errors	. Addresses	…

Figure 2. An example of F-term classification system.

map depicted in Figure 1 by selecting relevant dimensions from the possible viewpoints and by classifying patents based on the selected viewpoints.

Experts assign a patent to F-terms in two steps. They firstly determine themes of the patent, and then for each theme they assign the patent to F-terms. According to this procedure, we divided our subtask into two parts, "Theme Categorization Subtask" and "F-term Categorization Subtask". In Theme Categorization Subtask, participants determine one or more themes of each patent. This can be seen as a simplified version of classifying patents into the world standard taxonomy of IPC (International Patent Classification). Refer to [1] for approaches of automatic patent classification based on IPC[2]. In F-term Categorization Subtask, participants determine one or more F-terms of each patent whose theme has been given. F-term Categorization Subtask is a new attempt in that it is based on multi-dimensional (in other words, multifaceted) categories. In both subtasks, we provided huge number of training documents.

The rest of this paper is organized as follows. Section 2 introduces F-term classification system. Section 3 describes the designs of the two subtasks. Section 4 explains the datasets we released. Section 5 shows the evaluation results. Here we also introduce the participated systems briefly. Section 6 discusses the future directions. Section 7 concludes the paper.

2. F-term Classification System

The most common classification taxonomy of patents is IPC, which is internationally uniform. IPC is basically a single-dimension classification structure based solely on the contents of inventions. However, patent searchers sometimes have to explore patents focusing on various viewpoints such as purpose, function, effect, and so on. To this end, the Japan Patent Office provides a multi-dimensional classification structure called F-term classification system. Figure 2 shows an example.

In F-term classification system, each technological field is defined as a theme corresponding to a set of "FI" (an extension of IPC) codes. For example, the theme denoted by "5B001" is the technological field of "Detection and correction of errors (in computers)" and this theme corresponds to the FI codes of "G06F11/08-11/10,330@Z". A theme is expressed by a sequence of a digit, an alphabet, and three digits. There are over 2,500 themes.

Each theme has a collection of viewpoints for specifying possible aspects of the inventions under the theme[3]. For example, "5B001" has "PURPOSE", "MEANS" or "ERROR LOCATION" as viewpoints. The collection of viewpoints varies from theme to theme. In the example, "ERROR LOCATION" is appropriate only for this theme. A viewpoint is denoted by two alphabets. For example, "AC" represents the viewpoint "MEAN". Note that the naming policy of viewpoints is not uniform across themes, meaning that "AC" does not represent "MEAN" in other themes.

Each viewpoint has a list of possible elements. For example, "MEANS" in this theme can be "Code operations", "Comparison", "Interleaving", and so on. The collection of elements varies from viewpoint to viewpoint. An element is represented as two digits. For example, "Interleaving" for "MEAN" corresponds to "05". As an exception, "00" sometimes represents the elements not enumerated in the list of possible elements. The "00" element is also used to designate its belonging viewpoint, as seen in Figure 2.

A pair of a viewpoint and its element is shortly called F-term. For example, "AC05" is an F-term representing "mean (of error collection and correction) is interleaving". Although some F-terms

[2] Only class-level or subclass-level IPC categories (the numbers are 114 and 451 respectively) are considered in [1].

[3] Some themes do not have viewpoints mainly because FI is enough for classifying patents in these themes.

can have an additional alphabet for expressing more detailed classifications, we ignored the additional codes in this subtask.

There are general/specific relations between F-terms. This relationship is defined by dot (".") characters written in the description of each F-term. Figure 3 shows examples of descriptions.

3E003 (Container packaging and wrapping operations)

AA00 CONTAINERS
AA01 . Rigid containers
AA02 .. Rigid containers with integrated internal dividers
AA03 .. Rigid containers with separate internal dividers
AA04 .. Rigid containers with cushioning materials
AA05 . Soft containers

Figure 3. Examples of F-term descriptions.

The number of dots determines the level of hierarchy. No dot signifies the highest level, which is followed by single dot ("."), double dots (".."), and triple dots ("…") in the descending order of hierarchy. The F-terms in Figure 3 correspond to the hierarchy in Figure 4.

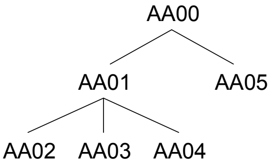

Figure 4. A hierarchy of F-terms.

In this subtask, we ignored hierarchy of F-terms because the evaluation becomes complicated when considering a partial match between F-terms. As a result, unless a participant submitted exactly the same F-term as the correct F-term, we regarded the assignment as wrong one even if both F-terms are in a general/specific relation.

3. Task Design

Figure 5 shows the overview of Classification Subtask. In this section, we explain the task designs of Theme Categorization Subtask and F-term Categorization Subtask. In the next section, we describe about the datasets in the figure.

3.1 Theme Categorization Subtask

In this subtask, participants had to submit a ranked list of 100 possible themes for each patent. Unlike the filtering track in TREC[4], our subtask is not for binary text classification where systems only have to decide for each document whether it should be accepted or rejected as a member of a category.

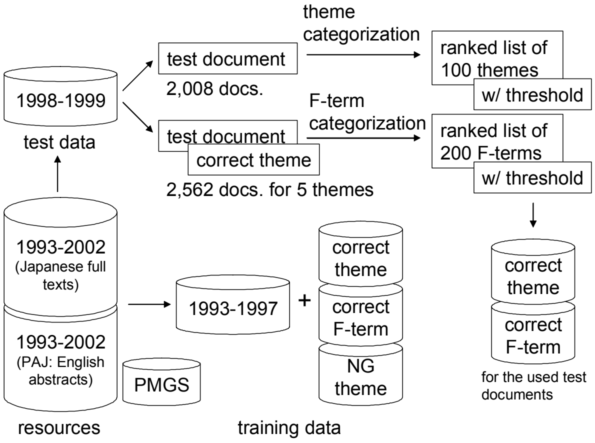

Figure 5. The overview of Classification Subtask.

In a ranked list, participants additionally had to determine the threshold of their confidence on theme assignments. The themes above the threshold were regarded as the ones submitted with confidence and those were used for calculating the F-measure.

Training documents of this subtask are full texts of Japanese patents published from 1993 to 1997, and test documents were randomly selected from those published from 1998 to 1999. Every Japanese full text has its English abstract and participants could use both collections in training and testing. The details of these datasets are described in Section 4.

Submitted results were evaluated based on recall/precision. For a ranked list for each test document, we calculated the 11 point interpolated precision, the MAP (Mean Average Precision), and the F-measure. These values were averaged over all the test documents (macro averaging).

Since almost all the test documents have only one or two themes (40% for one theme and 33% for two themes), interpolation of precision did not work effectively to distinguish recall/precision tradeoff curves between submitted results. For example, if a document has only one theme, the precision at the recall 0.0 is always interpolated by the precision at the recall 1.0, which means that the interpolated recall/precision curve becomes a horizontal line. If a document has two themes, the interpolated recall/precision curve becomes a shape of the two-step function (from 0.0 to 0.5 and from 0.5 to 1.0). In this subtask, since about 73% of the test documents have only one or two themes, the macro averaged recall/precision curve over the test documents has similar shape for every submitted result.

To address this problem, we additionally calculated the micro averaged precisions as follows. Assume that there are N test documents. We first collect K top-ranked categories for every test document and pool N*K categories. We then calculate the recall and the precision for this pool. For all values of K, we calculate the corresponding

[4] http://trec.nist.gov

Table 1. The themes used in F-term Categorization Subtask.

theme code	theme name	number of viewpoints	number of F-terms	examples of viewpoints
2B022	Cultivation of vegetables	9	95	"TARGET VEGETABLES", "MAIN COMPONENTS OF CULTURING MEDIA", "ENVIRONMENTAL CONTROL"
3G301	Electrical control of the air and fuel supply to internal combustion	21	369	"ENGINE MODELS", "GENERAL PURPOSE", "ENGINE TIMING CONTROL"
4B064	Manufacture of chemical compounds by using	23	541	"PRODUCTS CONTAINING OXYGEN", "SACCHARIDES AS THE PRODUCT", "MOLECULAR WEIGHT AS A PROPERTY"
5H180	Traffic-control systems	11	215	"OBJECTS TO BE CONTROLLED OR DETECTED", "MEANS OF DETECTION", "MANAGEMENT OF OPERATION OR TRAVELING OF INDIVIDUAL VEHICLES"
5J104	Ciphering device, decoding device and privacy communication	14	271	"purpose and effect", "form of telecommunication", "encryption method"

recall/precision values, which are used to interpolate the precisions at the 11 levels of recall.

3.2 F-term Categorization Subtask

In this subtask, participants had to submit a ranked list of 200 possible F-terms for each patent whose theme had been given. Participants also had to determine the threshold of their confidence on F-term assignments.

In this subtask, we used the five themes listed in Table1. Although the total number of possible F-terms across all the themes reaches to 337,027, the number of F-terms within each theme is relatively small. In this subtask, the numbers of possible F-terms for the five themes are between 95 and 541.

Training documents are full texts of Japanese patents published from 1993 to 1997 and test documents were randomly selected from those published from 1998 to 1999. English abstracts were allowed to use in training and testing.

Evaluation measures are basically the same as those in Theme Categorization Subtask. The only difference is that we do not need to calculate the micro averaged precisions. This is because interpolation of precision works effectively due to enough number of F-terms per test document (on average 11.4 F-terms).

4. Datasets

4.1 Document Resources

Unexamined Japanese patent applications published from 1993 to 2002 were released in this subtask. Those are full texts of Japanese patents (written in Japanese). The same years' English abstracts were also released. That is, every Japanese full text has its corresponding English abstract. This collection of English abstracts is called PAJ (Patent Abstract Japan).

At the same time, descriptions of themes and F-terms were released. This collection is called PMGS (Patent Map Guidance System)[5]. PMGS is provided in both Japanese and English.

4.2 Training data

For every patent published from 1993 to 1997, we released the lists of correct themes and correct F-terms as training data. Those themes and F-terms were taken from "Seirihyoujunka (Standardized) Data" which contains bibliographic information of patents in the SGML format. Seirihyoujukna Data was extracted from the master databases in the Japan Patent Office. Note that although some full texts include sections for their themes and F-terms, these themes and F-terms may not be the latest ones. In many times, themes and F-terms are added or deleted after publishing the texts and these revisions are reflected only on the databases in the Japan Patent Office.

4.3 Test data

In Theme Categorization Subtask, we randomly selected 2,008 patents from all the patents published from 1998 to 1999.

In F-term Categorization Subtask, we firstly selected five themes which have enough numbers of patents in every year and whose collections of viewpoints are typical ones. The five themes are listed in Table 1. For each theme, we randomly selected about 500 patents from the patents having the theme and published from 1998 to 1999.

4.4 NG Themes

We provided the list of NG (no-good) themes which were not used in this subtask. NG themes are the discontinued themes or the themes under revision. In addition to the NG themes, we did not use the theme "4K500" which is unofficially used in the Japan Patent Office

We did not filter out rare categories in constructing training/test data.

4.5 Distribution of categories

Table 2 and Table 3 show statistics of categories, and Figure 6 compares category distributions in training and test data. From these tables and figures, we can assert that the training and test data have similar distributions except for the theme "2B022" in F-term categorization. Here the average number of F-terms per test document (4.79) is twice as much as

[5] http://www5.ipdl.ncipi.go.jp/pmgs1/pmgs1/pmgs_E.

that per training document (2.76).

Table 2. Statistics of themes in Theme Categorization Task.

	number of documents	number of themes	average num of themes per docs
training	1667378	2519	2.03
test	2008	1445	2.04

Table 3. Statistics of F-terms in F-term Categorization Task.

		number of documents	number of F-terms	average num of F-terms per docs
2B022	training	1916	82	2.76
	test	475	73	4.79
3G301	training	6699	369	19.97
	test	542	353	21.13
4B064	training	6405	486	8.58
	test	493	323	8.49
5H180	training	6222	214	7.90
	test	508	174	9.21
5J104	training	1920	268	10.03
	test	544	251	12.07

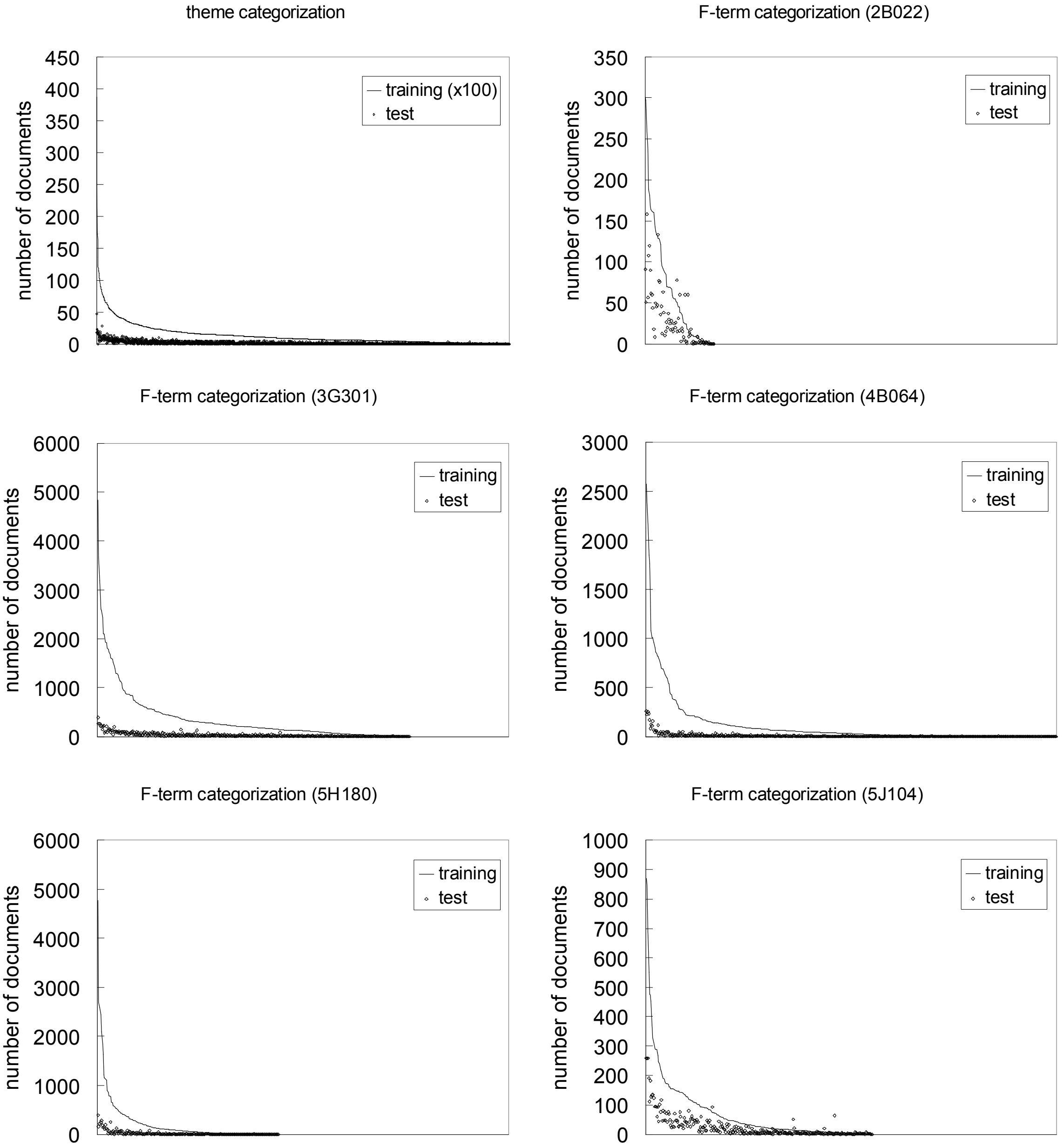

Figure 6. Distributions of themes and F-terms.

5. Evaluations

5.1 Results

We had five participating groups, four of them submitted results to Theme Categorization Subtask and three of them to F-term Categorization Subtask.

Figure 7 and Figure 9 are macro averaged recall/precision curves in the two subtasks. Table 4 and Table 5 compare the values of MAP and F-measure for the two subtasks. Only the best result (in MAP) from each participating group is on these figures and tables.

As shown in Figure 7, all the macro averaged recall/precision curves in theme categorization are almost the same shape because of few correct themes per patent. Figure 8 shows micro averaged version of recall/precision curves.

In theme categorization, K-NN was the best followed by Naive Bayes (shown in Table 4). In F-term categorization, K-NN was also the best (shown in Table 5).

Most of the participating groups constructed document surrogates by selecting informative components from each full text. Feature selection techniques were also used in some submissions. Two groups used PAJ in training and testing.

Figure 10 shows a theme-to-theme comparison in F-term categorization. Here we see the same ranking of the submitted results across the themes, although the MAP values are largely different from theme to theme.

By comparing the best MAP values between theme categorization (0.6872) and F-term categorization (0.4998), we can assume the difficulty of F-term categorization. Unfortunately, the participant who submitted the best result of theme categorization did not participate in F-term categorization, and vice versa. However, post experiments conducted by NICT[6] who had the best MAP of 0.4998 in F-term categorization shows that the MAP of theme categorization by the same approach is 0.6427, confirming the difficulty of F-term categorization.

5.2 Approaches

We briefly summarize the approaches of the submitted systems. For more information, refer to the original papers by the participants.

BOLA

BOLA participated in Theme Categorization Subtask. They submitted two approaches. The first one is based on K-NN, where the similarity between test and training documents is calculated based on structural similarity between them. Among components of Japanese patents, they use "technological field", "purpose", and "method". The second approach is based on Maximum Entropy, where two measures of term weighting were investigated for feature selection.

FXDM

FXDM participated in both subtasks. They applied their prototype system of document management to patent classification. Their approach is based on the vector space model, where each category is represented as a word vector and each test document is compared with these category vectors. Their experimental comparison of document surrogates showed that the combination of "technological field", "prior art", and "problems to be solved" is the best.

JSPAT

JSPAT participated in both subtasks. In theme categorization, they use a Naive Bayes approach with "shrinkage-based" probability estimation, where the probabilities under three layers' of conditions are estimated independently and linearly combined. In F-term categorization, they construct two kinds of binary SVMs for viewpoints and elements, and combine the results from these SVMs. Index terms they use are nouns, noun phrases, and sub-phrases of longer noun phrases.

NICT

NICT participated in F-term categorization subtask. Their approach is based on K-NN where retrieval model is BM25. They investigated three methods of category selection. For indexing, they select "abstract", "claim", "technological field", and "method" from Japanese patents. Although they did not participate in Theme Categorization Subtask, they conducted post experiments for theme categorization using the same approach used in F-term Categorization Subtask.

WGLAB

WGLAB participated in Theme Categorization Subtask. Their approach is based on K-NN, where retrieval model is BM11 or the vector space model. They investigated the following issues of categorization. The first is the number of K in K-NN; that is how many similar documents should be used. The second is the number of training documents in K-NN. The third is the issue on document surrogates. Their results showed that using PAJ is the best.

6. Future Directions

- We have to consider F-term hierarchy in F-term categorization, especially in evaluation. For example, when an F-term "f" is correct and "f" has "f1", "f2", and "f3" as the children, a system answering "f" should score higher then a system answering "f1", "f2" and "f3". What if a system answers "f" and "f2"? It is needed a consistent

[6] For more detail, refer to the paper by NICT.

measure of evaluating partial matches between F-terms.

- We have to evaluate more themes in F-term categorization. In this subtask, we selected only five themes that have typical viewpoints like purpose, method, etc. In the selection process, we excluded the F-terms categorizing to those needs numerical analysis. For example, since the theme “5F045” (“Vapor-phase growth”) has the F-term “AD05” standing for “the growth condition is the temperature T where 100C <= T < 200C”, we did not use “5F045” as our candidate themes.
- F-term assignment is more similar to indexing than to categorization. In fact, when experts assign a patent to F-terms, they read the patent and mark related words or phrases to the F-terms. In F-term categorization subtask however, we could not have approaches based on indexing. In future we have to compare the indexing approach to the categorization approach. For this purpose, we are planning to digitize experts’ annotations about F-term assignments.
- We have to evaluate the use of F-terms in patent map generation which is the ultimate goal of our project.

7. Conclusion

In Classification Subtask at NTCIR-5 Patent Retrieval Task, we released test collections for patent classification. The test collections are based on F-term classification system which has multi-dimensional category structure. Using the test collections, we performed two subtasks for theme categorization and F-term categorization; the former is a common text categorization and the latter is a text categorization based on multifaceted categories.

References

[1] C.J. Fall, A. Torcsvari, K. Benzineb, and G. Karetka. Automated Categorization in the International Patent Classification. ACM SIGIR Forum, Vol.37, No.1, pp.10-25, 2003.

[2] A. Fujii, M. Iwayama, and N. Kando. Overview of Patent Retrieval Task at NTCIR-4. In Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies, Information Retrieval, Question Answering and Summarization, 2004.

[3] I. Schellner, Japanese File Index classification and F-terms. World Patent Information, Vol.24, pp.197-201, 2002.

Table 4. Results of Theme Categorization Subtask.

Runid	model	MAP	R-Precision	F-measure
BOLA1	K-NN	0.6872	0.5943	0.2690
JSPAT2	Naive Bayes	0.6591	0.5634	0.5269
WGLAB9	K-NN	0.6192	0.5305	0.0682
FXDM3	VSM	0.4886	0.3881	0.3778

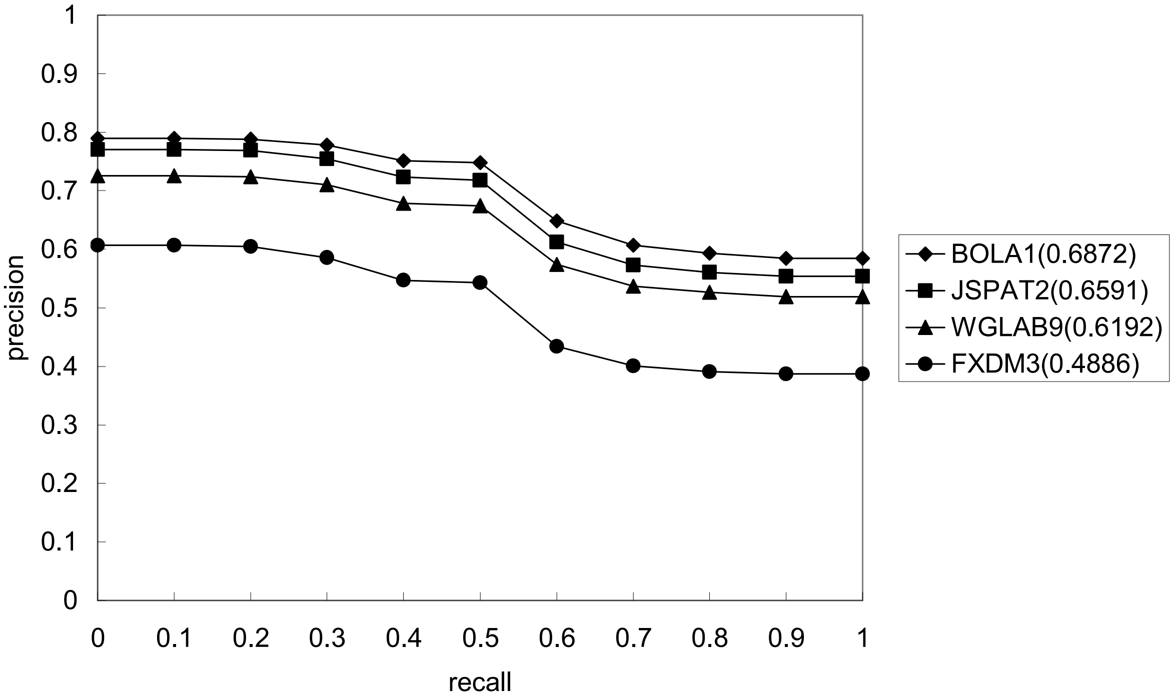

Figure 7. Macro averaged recall/precision curves for Theme Categorization Subtask.

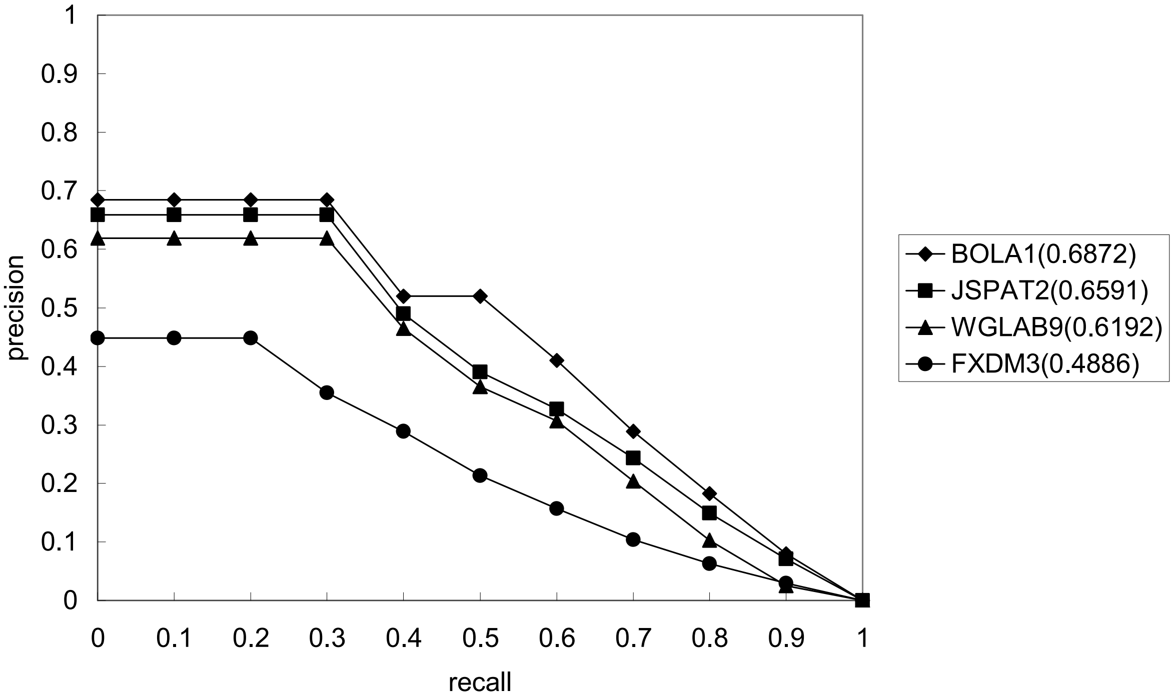

Figure 8. Micro averaged recall/precision curves for Theme Categorization Subtask.

Table 5. Results of F-term Categorization Subtask.

Runid	model	MAP	R-Precision	F-measure
NICT5	K-NN	0.4998	0.4611	0.4393
JSPAT1	SVM	0.3990	0.3879	0.2830
FXDM10	VSM	0.2052	0.1989	0.1579

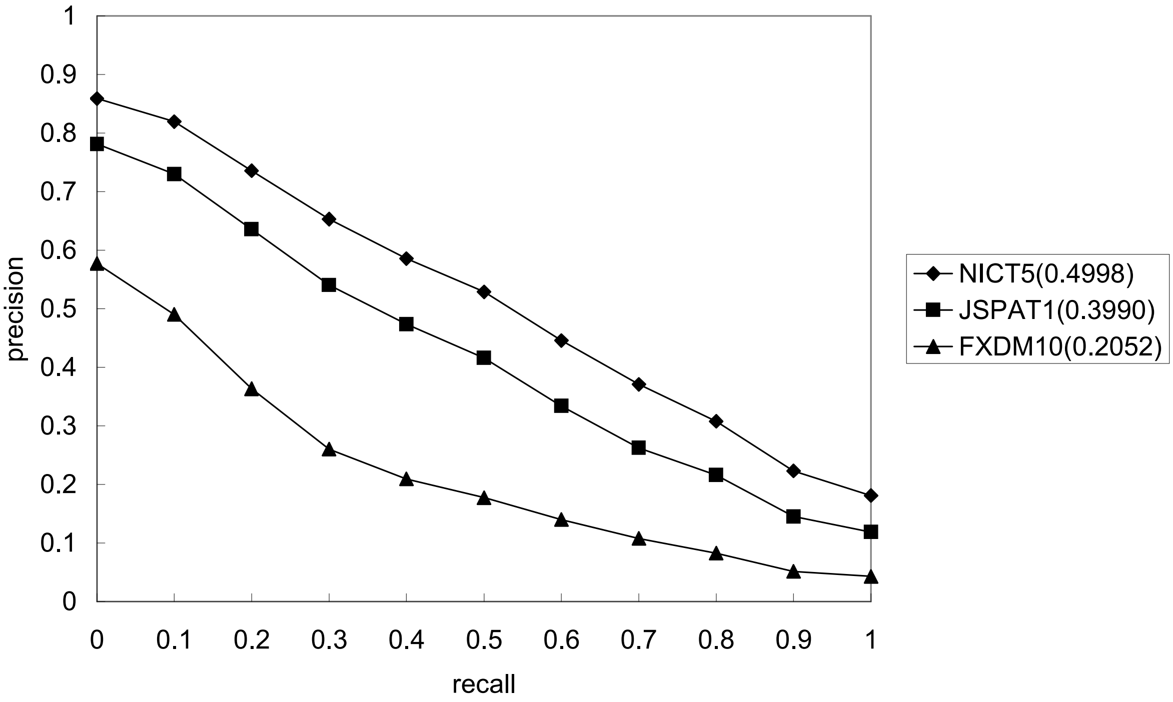

Figure 9. Macro averaged recall/precision curves for F-term Categorization Subtask.

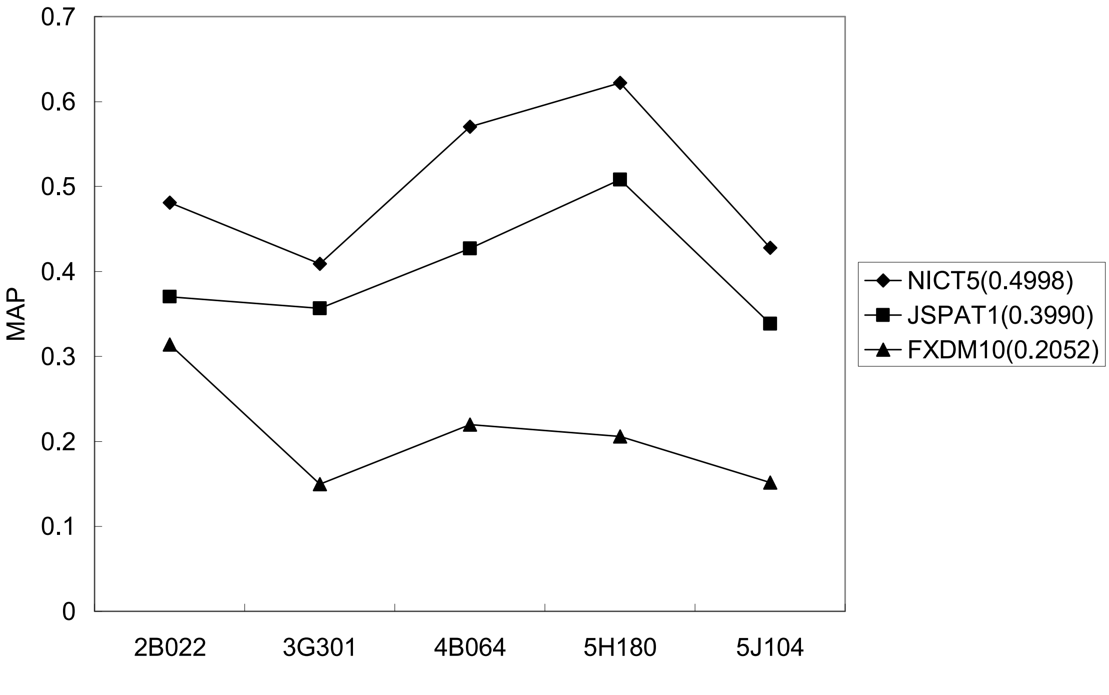

Figure 10. Theme-to-theme comparison of MAPs in F-term categorization Subtask.