Memorandum of the first round-table meeting for the NTCIR-7 Patent Mining Task

Edited by Hidetsugu Nanba

Time & Date:

10:00-12:00, March 31, 2008. (Mon.)

Place:

room #1213 at National Institute of Informatics

(2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 JAPAN)

Participants: (alphabetical order)

Makoto Iwayama (Hitachi, Ltd.) (organizer)

Noriko Kando (National Institute of Informatics)

Hisao Mase (Hitachi, Ltd.)

Hidetsugu Nanba (Hiroshima City University) (organizer)

Takayuki Shimano (Nagaoka University of Technology)

Takashi Yukawa (Nagaoka University of Technology)

Agenda:

(1) Report of the dry run

Refer to files "200803_round_table_e.pdf" and "overview.pdf".

(2) Discussion

- **[Comment 1]**

  How about showing a list of organizations of participants to all participant groups? It may be useful to guess other participants' approaches from the list.

- **[Question 1]**

  The information in abstracts (topics) is not much enough. Isn't it possible to use full text data as topics, even though the number of the texts is not so much?

  [Answer from the organizers]

  We also think that using full texts is ideal. However, it is difficult to use in NTCIR-7.

- **[Question 2]**

  The number of IPC codes in pseudo-training data is smaller than those in the sixth

edition of IPC system. Actually, some IPC codes were not extracted from patents, even though they were in the sixth edition.

[Answer from the organizers]

The organizers preliminarily removed some IPC codes, which do not relate to academic fields, from a list of IPC distributed at the dry run web page. In the following, we will briefly explain the procedure. As we described in Section 2.2 in overview (please refer to "overview.pdf"), we focused on "Indication of exceptions to lack of novelty" field (exception field) in Japanese patents to create the data for evaluation. Now, we also use this exception field to select IPC codes relating to academic fields. Generally, patents containing exception fields seem to be related with academic fields. Therefore, we extracted IPC codes from such patents, and made the list of IPC.

- **[Question 3]**

  Are there any difference between topics 100-151 and topics 200-244? Among topics 200-244, there were five topics that our system could not detect correct IPC codes within top 1000 results.

  [Answer from the organizers]

  IPC codes for topics 100-151 are more relevant than those for topics 200-244. In the procedure of creating data for evaluation, which we described in Section 2.2 in the overview, we identified two research papers (a paper in an exception field and a record in a research paper database of NTCIR-1 and 2) based on the following two ranks:

  - Two papers are exactly the same (topics 100-151)
  - Authors and research topics of two papers are almost the same, but the publication years are different (topics 200-244)

- **[Comment 2]**

  Even though more than one IPC codes are assigned to a patent, the first one, called Hit IPC, listed in a patent is more important than others. It is better to evaluate how well each system could detect Hit IPCs.

- **[Comment 3]**

  There are a lot of typos in topics, which degraded the system's performance.

- **[Report from HTC group]**

  HTC group submitted four systems, HTC01, HTC02, HTC03, and HTC04, to the dry run Japanese subtask. All these systems identified IPC codes based on the KNN method using the results of Patent retrieval system. Only the difference among these systems is the number of patents retrieved by the IR system. The results are shown in Table 1 (refer to the 17th slide in "200803_round_table_e.pdf").

Table 1. The results of four systems of the HTC group

	Top n	MAP
HTC01	1	0.6764
HTC02	5	0.6293
HTC03	10	0.5585
HTC04	1,000	0.3635

As can be seen from Table 1, HTC01 was the best. Empirically, the HTC group's system obtained the best score for other classification task, when the system used between top 15 to 50 results. The main reason for these results is that the system seems to detect patents, which are the counterparts of research papers (topics) at top one result.

To degrade the effect of the counterparts, we removed top one from a list of retrieved patents, and applied the KNN method to the list again. As a result, the system obtained 0.2458 of a MAP score, when we used around top 15 and 50 results of our IR system, which matched their empirical results. Therefore, it seems better to remove patents, which are the counterparts of research papers (topics) from the patent corpus.

[Answer from the organizers]

We will employ this idea to the formal run. We will release modified version of pseudo-training data.

- **[Comment 4]**

  The number of patents in each IPC code is different every year. How about evaluating systems by publication years of research papers (topics)?