Overview of the Topical Classification Task at NTCIR-4 WEB

Koji Eguchi †

† National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
eguchi@nii.ac.jp

Abstract

This paper gives an overview of the Topical Classification Task 1 that was conducted from 2003 to 2004 as a subtask of the WEB Task at the Fourth NTCIR Workshop ('NTCIR-4 WEB'). In this Topical Classification Task, we attempted to assess the effectiveness of automatic classification systems for retrieved documents from Web search engine systems from a viewpoint of topical relevance. For the classification task we used a target data set comprising ranked lists of search result documents from 100-gigabyte document data, which were mainly gathered from the '.jp' domain. We carried out an evaluation of classification systems on the basis of the information retrieval task. We applied several evaluation measures that are often used in information retrieval evaluation. We also proposed new evaluation measures considering the number of classes. We carried out an evaluation considering duplicate documents that is suitable for the evaluation of non-exclusive classification systems.
Keywords: *Text Classification, Document Clustering, Evaluation Methods.*

1 Introduction

This paper gives an overview of the Topical Classification Task 1 that was conducted from 2003 to 2004 as a subtask of the WEB Task at the Fourth NTCIR Workshop ('NTCIR-4 WEB'). This Topical Classification Task addressed evaluation of techniques that classifies search results for a user input query to relax the user's cognitive loads. In such kind of cases, classification viewpoint can be topics of web pages, genres or types, layouts, author or institutional information given by URL strings, space distribution of geographic features described in web pages, last-updated time or order relation of events described in web pages, and so on. At the NTCIR-4 WEB, we focused on the web page classification task based on topics.

2 Overview of the Task

2.1 Task Description

We conducted a dry run and a formal run. The dry run aimed at sharing common views between participating groups and organizers, and clarifying problems in system design by the participating groups and in evaluation methods by the organizers to reflect them toward the formal run. In the dry run, we prioritize gaining experience in the procedure for this task in the shortest possible time with small-scaled experiments, rather than with plenty of scale for such as parameter tuning. In this paper, we describe on the formal run, as long as we do not have to specify as the dry run.

Participating groups classified a part of target data set comprising ranked lists of search result documents for search requests, as being described in Section **??**. They only classified up to the N-th documents of the target data, not entire target data, into an adequate number of classes, *i.e.*, categories or clusters, for each search request. Here, we set the value of N as $N = 200$. Then, they submitted the classification results to organizers. We did not fix a class structure in advance[1]. We also did not limit the number of documents to be included in individual classes. It is desirable that documents classified into each class are ranked in adequate order within the class. Each class had to have a class label. A machine-like identification code was allowed to be used as the class label, however, more readable labels such as topical terms, typical page titles or short summary sentences or snippets were more desirable. The participating groups submitted not only classification results but also 'system description forms' in the proper form to the organizers.

[1]We explored the possibility of imposing limitations of the number of classes, by carrying out a questionnaire survey to the participating groups. As the results, the number of the participating groups who requested not to impose the limitations was larger than the ones who requested to do so. So, we decided not to impose the limitations of the number of classes as previously mentioned.

2.2 Envisioned Technical Approaches

Various approaches can be viewed as means of web page classification task, such as document clustering based on contents of documents, clustering based on hyperlink connections, clustering based on term co-occurrence within particular tags, text categorization based on a known subject/classification structure, combination of some of these techniques. The organizers did not deliver training data or a subject/classification system. We used possible general methods for evaluation, rather than methods specialized to an individual approach.

We did not prevent the participating groups from using approaches that involves interactions with users on the process of classifying documents. In such kind of cases, they are expected to submit only the final classification results to be evaluated by the organizers, and to specify the existence of the human intervention in the system description forms. In comparative evaluation, the organizers planned to divide the classification results that involve human intervention from the others, so that the both of types of the results were separately evaluated. As the result, however, all the classification results were obtained automatically by the participating groups.

We also did not prevent the participating groups from using approaches that assigns a document to several classes. In this case, they were also expected to specify the use of the approach in the system description forms.

When participating groups used a hierarchical classification approach, they were expected to submit the classification results where each document is assigned to a non-hierarchical structure, selecting one of the layers of hierarchy.

3 Data for Classification Processing

3.1 Topics and Target Data Set

We used a part of the topics that were created in the Informational Retrieval Task at the NTCIR-4 WEB. All of the topics were written in Japanese. A topic example and its English translation are shown in **Figure 1**.

- ⟨TOPIC⟩ specified the boundary of a topic.
- ⟨NUM⟩ indicated the topic identification number.
- ⟨TITLE⟩ gives 1-3 terms that are simulated by the topic creator to be similar to query terms used in real Web search engines. The terms in the ⟨TITLE⟩ are listed in their order of importance for searching. The ⟨TITLE⟩ has the attribute of 'CASE' that indicates the types of search strategies, as follows:

⟨TOPIC⟩
⟨NUM⟩0001⟨/NUM⟩
⟨TITLE CASE="c" RELAT="2-3"⟩ オフサイド, サッカー, ルール ⟨/TITLE⟩
⟨DESC⟩ サッカーのオフサイドというルールについて説明されている文書を探したい ⟨/DESC⟩
⟨NARR⟩⟨BACK⟩ サッカーでオフサイドとはどういうルールなのかを知りたい。⟨/BACK⟩⟨TERM⟩ オフサイドはオフェンス側の反則である。オフサイドが適用される状況にはいくつかのパターンがあり、サッカーのルールの中で最もわかりにくいものである。⟨/TERM⟩⟨RELE⟩ 適合文書はオフサイドが適用される状況を説明しているもの ⟨/RELE⟩⟨/NARR⟩
⟨ALT0 CASE="b"⟩ オフサイド ⟨/ALT0⟩
⟨ALT1 CASE="b"⟩ オフサイド, 選手, 位置 ⟨/ALT1⟩
⟨ALT2 CASE="b"⟩ オフサイド, サッカー ⟨/ALT2⟩
⟨ALT3 CASE="b"⟩ サッカー, オフサイド, ルール ⟨/ALT3⟩
⟨USER⟩ 大学 2 年, 男性, 検索歴 4 年, 熟練度 3, 精通度 5⟨/USER⟩
⟨/TOPIC⟩

(a) An original sample topic

⟨TOPIC⟩
⟨NUM⟩0001⟨/NUM⟩
⟨TITLE CASE="c" RELAT="2-3"⟩offside, soccer, rule⟨/TITLE⟩
⟨DESC⟩ I want to find documents that explain the offside rule in soccer. ⟨/DESC⟩
⟨NARR⟩⟨BACK⟩ I want to know about the offside rule in soccer. ⟨/BACK⟩⟨TERM⟩ Offside is a foul committed by a member of the offense side. There are several patterns for situations in which the offside rule can be applied, and it is the most difficult soccer rule to understand. ⟨/TERM⟩⟨RELE⟩ Relevant documents must explain situations where the offside rule applies. ⟨/RELE⟩⟨/NARR⟩
⟨ALT0 CASE="b"⟩offside⟨/ALT0⟩
⟨ALT1 CASE="b"⟩offside, player, position⟨/ALT1⟩
⟨ALT2 CASE="b"⟩offside, soccer⟨/ALT2⟩
⟨ALT3 CASE="b"⟩soccer, offside, rule⟨/ALT3⟩
⟨USER⟩2nd year undergraduate student, male, 4 years of search experience, skill level 3, familiarity level 5⟨/USER⟩
⟨/TOPIC⟩

(b) An English translation of a sample topic

Figure 1. A sample topic and its English translation

  (a) All of the terms have the relation one another that can be used as OR operator.
  (b) All of the terms have the relation one another that can be used as AND operator.
  (c) Only two of the terms have the relation that can be used as OR operator, and are specified by the attribute of 'RELAT'.

- ⟨ALT0⟩ was created as the result of extracting the first appeared term in the ⟨TITLE⟩ field of the topic. The ⟨ALT0⟩ field has the term judged as being most important for searching, since the terms in the ⟨TITLE⟩ field were listed in the order of importance for searching.

Other details of the topic data can be found in Reference [4].

The topic data were composed of 47 topics at the head of 153 topic data of the Informational Retrieval

Task, and were delivered to the participating groups. However, we used only a part of the 47 topic data for evaluation. The topic selection strategies we used are as follows:

(1) We selected the topics whose relevant documents are neither too many nor too few, after assessing relevance of the documents to search requests.

(2) We selected the topics that could be judged as being comparatively more ambiguous or broader.

The organizers selected five search result lists that were retrieved from the 100-gigabyte 'NW100G-01' using only the ⟨ALT0⟩ part in each of the topics of the Informational Retrieval Task. The ⟨ALT0⟩ tag specified one query term. Then, the organizers merged up to 100 documents in each search result list into one using a meta-search-engine method, the 'Borda Count' voting algorithm [1]. The resulting list is used as 'target data', which was delivered to the participating groups along with the topic data.

3.2 Other Available Data Provided by Organizers

Although the rest of the target data set ranked after the $(N+1)$-th document could not be used as being classified, they could be used on the process of classifying documents. 'NW100G-01,' [6, 5] the 100-gigabyte document data set that were used for searching could also be used on the process of classifying documents. The NW100G-01 included hyperlink relationship data as well as web page data. Any parts of the topic statement could be used on the process of classifying documents.

The participating groups were expected to specify which kinds of data as mentioned above were used on the process of classifying documents as the system description forms, and to submit them together with the classification results to the organizers.

3.3 External Resources

If the participating groups who had any plans to use alternative data resources ('external resources') that were available for the public or were held by their own, other than the available data as mentioned in Section **3.2**, they were expected to communicate with the organizers prior to carrying out the classification task because some kinds of the external resources seemed not to be proper to the comparative evaluation. As the result, no participating groups made a declaration of using the external resources.

4 Evaluation Methods

While assessing the results, we assume the following two stages: (i) where the user possesses explicit search requests in his/her mind in browsing search results, and (ii) where the user does not possess explicit search requests in searching and in browsing the search results.

We attempted to carry out two types of evaluation, as being described in Sections **4.2** and **4.3**, assuming previously mentioned (i) and (ii), respectively. First of all, we describe the relevance judgment data in Section **4.1**, which were partially used for evaluation.

4.1 Relevance Judgment Data

At the Informational Retrieval Task at the NTCIR-4 WEB, the assessors judged the 'multi-grade relevance' of the individual documents as: highly relevant, fairly relevant, partially relevant or irrelevant. Here, the number of documents corresponding to each grade was not controlled —for example, the assessor did not care if the number of highly relevant documents were very small—.

The assessors judged the relevance of the documents only on the basis of the information given in Japanese or English. The documents included in the document data seemed to be described in various languages, because we had not discarded documents with page data described in languages other than Japanese or English from the document data. If a part of the documents were entirely described in languages other than Japanese or English, the assessors must have judge this kind of documents as being irrelevant.

The assessor judged the relevance of a page when he/she could browse the page and its out-linked pages that satisfied some of the conditions, but not all of the out-linked pages. The out-linked pages indicate pages that are connected from a certain page whose anchor tags describe the URLs of the out-linked pages [4].

4.2 Evaluation based on Distribution of Relevant Documents

The evaluation described in this section assumes the stage where the user possesses explicit search requests in his/her mind while browsing search results. In this evaluation we computed numerical indices based on location of relevant documents, such as precision and recall, using document ranking within classes with largest number of relevant documents [7, 10, 3]. We used relevance judgment data of the Informational Retrieval Task at the NTCIR-4 WEB, as described in Section **4.1**.

We sorted the classes in order of the number of relevant documents included, then extracted highly ranked n documents from the highly ranked classes. In this paper, we set the value of n as $n = 20$. If the top ranked class included more than n documents, we cut off the documents ranked lower than the n-th documents. If the top ranked class included less than n documents, we used the n documents gathered from several of the highly ranked classes.

Using the n documents obtained as above, we evaluate classification systems using some evaluation measures. Several of them are often used in information retrieval evaluation, as being described in Sections **4.2.1** and **4.2.2**. The other measures are proposed in these experiments as being described in Section **4.2.3**. We summarize these evaluation measures in Section **4.2.4**. We also carried out an evaluation considering duplicate documents that can be used for the evaluation of non-exclusive classification systems, as being described in Section **4.2.5**.

4.2.1 Evaluation Measures based on Precision and Recall

We calculated the 'average precision (non-interpolated)' measure. We also calculated the precision and recall for the n documents, and using them, calculated the 'F-value,' [2] which is the harmonic mean of the recall and precision as

$$\mathrm{f}(n) = \frac{2}{\frac{1}{\mathrm{p}(n)} + \frac{1}{\mathrm{r}(n)}}. \tag{1}$$

where $\mathrm{f}(n)$, $\mathrm{p}(n)$ and $\mathrm{r}(n)$ indicate the F-value, precision and recall for the n documents, respectively.

We define the following two relevance levels to calculate these measures.

(RL_1) Rigid relevance level For the precision-recall-related measures at the Rigid relevance level, we considered the document to be relevant if it was highly relevant or fairly relevant, and otherwise considered it to be irrelevant.

(RL_2) Relaxed relevance level For the precision-recall-related measures at the Relaxed relevance level, we considered the document to be relevant if it was highly relevant, fairly relevant or partially relevant, and otherwise considered it to be irrelevant.

4.2.2 Discounted Cumulative Gain

We adopted 'Cumulative Gain' ('CG') and 'Discounted Cumulative Gain' ('DCG') measures [8, 9] as one of the evaluation measures suitable for multi-grade relevance. The CG and DCG are represented by the following equations:

$$\mathrm{cg}(i) = \begin{cases} \mathrm{g}(1) & \text{if } i = 1 \\ \mathrm{cg}(i-1) + \mathrm{g}(i) & \text{otherwise}, \end{cases} \tag{2}$$

$$\mathrm{dcg}(i) = \begin{cases} \mathrm{g}(1) & \text{if } i = 1 \\ \mathrm{dcg}(i-1) + \mathrm{g}(i)/\log_b(i+1) & \text{otherwise}, \end{cases} \tag{3}$$

where

$$\mathrm{g}(i) = \begin{cases} h & \text{if } \mathrm{d}(i) \in H \\ a & \text{if } \mathrm{d}(i) \in A \\ b & \text{if } \mathrm{d}(i) \in B \end{cases} \tag{4}$$

where $\mathrm{d}(i)$ indicates the i-th-ranked document, and H, A and B indicate the sets of highly relevant, fairly relevant, and partially relevant documents, respectively. We set the magnitude of the gain indicated in Equation (4) to the following two relevance levels:

(RL_1) Rigid relevance level
$(h, a, b) = (3, 2, 0)$,

(RL_2) Relaxed relevance level
$(h, a, b) = (3, 2, 1)$.

We set the base of the logarithmic function as $b = 2$ in Equation (3). The DCG was derived from the CG, and modified in that the gain $\mathrm{g}(i)$ at rank i was discounted as being divided by a logarithmic rank i [8].

4.2.3 Discounted Cumulative Gain with Modification

We modified the DCG to evaluate classification results in the following two ways.

$$\mathrm{mdcg}_1(i) = \begin{cases} \mathrm{g}(1) & \text{if } i = 1 \\ \mathrm{mdcg}_1(i-1) + \mathrm{g}(i)/\{\log_b(i+1) & \\ \qquad \times \log_b(k(j+1))\} & \text{otherwise}, \end{cases} \tag{5}$$

$$\mathrm{mdcg}_2(i) = \begin{cases} \mathrm{g}(1) & \text{if } i = 1 \\ \mathrm{mdcg}_2(i-1) + \mathrm{g}(i)/\log_b(j+1) & \text{otherwise}, \end{cases} \tag{6}$$

where i indicates the ranking of the document, as in Equation (3), and j indicates the ranking of the class where the i document is included. For simplicity, we set k as $k = 1$ in Equation (5). We set the magnitude of the gain $g(i)$ as indicated in Equation (4).

We call these measures as 'MDCG' (Discounted Cumulative Gain with Modification) in this paper.

4.2.4 Characteristics of Evaluation Measures

Using mdcg_1 and mdcg_2 measures given by Equations (5) and (6), the lower a class is ranked, the more

$g(i)$ of a document included in the class becomes to be discounted in the case that the classification result has a large number classes. Therefore, this measure can be said to as evaluation measures considering the number of classes.

cg and mdcg_2 measures given by Equations (2) and (6), respectively, are independent from the document ranking within classes. So, those can be used to the classification systems without document ranking.

On the contrary, dcg and mdcg_1 measures given by Equations (3) and (5), respectively, depend on the document ranking within classes. So, those can be used to the classification systems without document ranking. So, those can be used to the classification systems with document ranking.

4.2.5 An Evaluation Considering Non-exclusive Classification

In the case of using non-exclusive classification systems, duplicate documents having the same document identification number are possible to appear in the n documents used for evaluation that were described at the second paragraph of Section **4.2**. We carried out an evaluation considering such cases as described below.

- For the document, comprising the duplicate document group, that first appeared in each run result list, we treated this kind of document as it is.
- For the other duplicate documents, we treated them as irrelevant although they were judged as relevant.

Consequently, non-exclusive classification results that contained many duplicate documents were expected to pay a penalty.

We designed this evaluation method by supposing it to be combined with the precision-recall-related measures, the DCG measure or its variations, which were described in Sections **4.2.1**, **4.2.2** and **4.2.3**, respectively.

4.3 Intrinsic Evaluation based on Classification Relevance

The evaluation in this section assumes the stage where the user does not possess explicit search requests while searching and browsing the search results. The evaluation method is basically based on relevance of documents included in each class to a typical topic of the class, while the method described in Section **4.2** is based on relevance to the topic of the search request.

We describe the evaluation of document classification into the class structure. An assessor browsed the classification results to understand overview of each class using contents of included documents. Then, he/she found a document that was classified in an inadequate class, and specified another class where he/she judged as being most adequate for the document. The assessor did not amend the class structure of the classification result in principle. The organizers evaluated the classification result using the assessor's judgment results.

We tried to carry out this kind of evaluation as the dry run; however, it required a large amount of assessment costs and time. We would like to describe the result of this evaluation later.

4.4 Other Evaluation Methods

We describe other evaluation methods from various viewpoints, although we have not used such methods, as follows:

(a) Evaluation of class structure. This can be based on 'understandability' of the class structure in the classification result.

(b) Comparative evaluation using reference data created by the users. To create to reference data, the assessor classifies up to N-th documents of the target data set, not being provided classification results. The class structure of each classification result will be compared with the one of the reference data.

(c) Appropriateness evaluation of labels on the classes.

(d) Comparative evaluation oriented for an individual approach.

5 Evaluation Results

5.1 Summary of Participation

Five groups, listed below in alphabetical order of affiliations, submitted their completed run results.

- Ibaraki University
- Matsushita Electric Industrial Co. Ltd.
- NTT Communication Science Laboratories
- Tokyo Metropolitan University
- Toyohashi University of Technology

The individual participating groups pursued various objectives. We summarize them in **Figure 2**, which are derived from the system description forms that were submitted by participating groups. Details of the system description forms are described in **Appendix**. Further details can be found in papers of the participating groups in this volume.

Table 1. Evaluation results

	RunID	Avg.Prec.	p(20)	r(20)	f(20)	cg(20)	dcg(20)	$mdcg_1(20)$	$mdcg_2(20)$
Rigid	ELRG-01	0.2454	0.2727	0.3405	0.1997	5.4545	2.2193	2.2193	5.4545
	METAL-01	0.3604	0.4455	0.7496	0.3725	8.6364	3.3282	3.1787	8.1463
	METAL-02	0.3579	0.4500	0.7502	0.3735	8.7273	3.2843	3.1155	8.1582
	METAL-03	0.3604	0.4455	0.7496	0.3725	8.6364	3.3282	3.1787	8.1463
	METAL-04	0.3620	0.4545	0.7684	0.3808	8.8182	3.3020	3.1103	8.1598
	SRLAB-01	0.2167	0.3455	0.5253	0.2701	6.9091	2.6490	2.4546	6.1450
	iba-02	0.0998	0.1773	0.3202	0.1431	3.5455	1.2971	1.2885	3.5119
	smlab-01	0.0586	0.2500	0.1682	0.1414	5.0000	1.6202	1.6202	5.0000
	avg.	0.2576	0.3551	0.5465	0.2817	6.9659	2.6285	2.5207	6.5902
Relaxed	ELRG-01	0.1279	0.3409	0.2317	0.2126	6.8182	2.5710	2.5710	6.8182
	METAL-01	0.3011	0.4773	0.5189	0.3495	9.1818	3.6512	3.5375	8.8008
	METAL-02	0.2987	0.4818	0.5385	0.3540	9.3636	3.7184	3.5835	8.8792
	METAL-03	0.3011	0.4773	0.5189	0.3495	9.1818	3.6512	3.5375	8.8008
	METAL-04	0.2998	0.4818	0.5535	0.3575	9.4545	3.7604	3.6014	8.8990
	SRLAB-01	0.2090	0.4045	0.4838	0.3044	8.0909	3.0121	2.7989	7.2478
	iba-02	0.1179	0.2318	0.2908	0.1857	4.6364	1.6555	1.6282	4.5574
	smlab-01	0.0621	0.2818	0.1599	0.1523	5.6364	1.9072	1.9072	5.6364
	avg.	0.2147	0.3972	0.4120	0.2832	7.7955	2.9909	2.8957	7.4549

5.2 Summary of Evaluation Results

Selected evaluation results of the Topical Classification Task at NTCIR-4 WEB are shown in **Table 1**. The evaluation was carried out at both the Rigid relevance level RL_1 and the Relaxed relevance level RL_2. All the evaluation was performed considering Non-exclusive Classification, as described in Section **4.2.5**. All the evaluation values were averaged over 11 topics. We will evaluate for larger number of topics later.

6 Conclusions

We carried out an evaluation of classification systems on the basis of the information retrieval task. We applied several evaluation measures that are often used in information retrieval evaluation. We also proposed new evaluation measures considering the number of classes. An intrinsic evaluation based on classification relevance is currently in progress. We carried out an evaluation considering duplicate documents that is suitable for the evaluation of non-exclusive classification systems. Detailed analysis of the evaluation results is an issue in the nearest future.

Acknowledgements

This work was partially supported by the Grants-in-Aid for Scientific Research on Priority Areas of "Informatics" (#13224087) and for Encouragement of Young Scientists (#14780339) from the Ministry of Education, Culture, Sports, Science and Technology, Japan. We greatly appreciate the efforts of all the participants of the Topical Classification Task 1 of the WEB Task at the Fourth NTCIR Workshop. We would like thank Dr. Akiko Aizawa at National Institute of Informatics for her useful advice.

References

[1] J. A. Aslam and M. Montague. Models for metasearch. In *Proceedings of the 24th Annual International ACM SIGIR Conference*, pages 276–284, New Orleans, Louisiana, USA, Sep. 2001.

[2] R. Baeza-Yates, editor. *Modern Information Retrieval*. Addison-Wesley, 1999.

[3] K. Eguchi, H. Ito, A. Kumamoto, and Y. Kanata. Adaptive document clustering using incrementally expanded queries. *Systems and Computers in Japan*, 32(2):64–74, Feb. 2001.

[4] K. Eguchi, K. Oyama, A. Aizawa, and H. Ishikawa. Overview of the informational retrieval task at NTCIR-4 WEB. In *Working Notes of the 4th NTCIR Workshop Meeting*, Tokyo, Japan, Jun. 2004.

[5] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama. Evaluation methods for web retrieval tasks considering hyperlink structure. *IEICE Transactions on Information and Systems*, E86-D(9):1804–1813, Sep. 2003.

[6] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama. Overview of the Web Retrieval Task at the Third NTCIR Workshop. Technical Report NII-2003-002E, National Institute of Informatics, Jan. 2003.

[7] M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. In *Proceedings of the 19th Annual International ACM SIGIR Conference*, pages 76–84, 1996.

[8] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference*, pages 41–48, Athens, Greece, Jul. 2000.

[9] E. M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference*, pages 74–82, Sep. 2001.

[10] O. Zamir and O. Etzioni. Web document clustering: A feasibility demonstration. In *Proceedings of the 21st Annual International ACM SIGIR Conference*, pages 46–54, 1998.

RunID	ELRG-01	METAL-01	METAL-02	METAL-03	METAL-04	SRLAB-01	iba-02	smlab-01
Subtask	D	D	D	D	D	D	D	D
Method	automatic	automatic	automatic	automatic	automatic	automatic	automatic	automatic
ClsModel	Document clustering based on term co-occurrence.	content-based clustering utilizing extracted feature terms	content-based clustering utilizing extracted feature terms	content-based clustering utilizing extracted feature terms	content-based clustering utilizing extracted feature terms	bag of words model with a hierachical PLSA with BIC	document clustering based on contents of documents	document classification based on suitability for specialists
LinkInfo	none	none	none	none	none	none	none	none
QueryInfo	none	used for excluding query terms from feature terms	used for excluding query terms from feature terms	used for excluding query terms from feature terms	used for excluding query terms from feature terms	none	none	none
Exclusion	exclusive classification	non-exclusive classification	non-exclusive classification	non-exclusive classification	non-exclusive classification	exclusive classification	exclusive classification	exclusive classification
Hierarchy	non-hierarchical classification	hierarchical classification, but classes in results file are top level classes only.	hierarchical classification, but classes in results file are top level classes only.	hierarchical classification, but classes in results file are top level classes only.	hierarchical classification, but classes in results file are top level classes only.	hierarchical classification	hierarchical clustering	non-hierarchical classification
Ranking	given score of meta search engine	ranking information of original meta search results is preserved in each class	ranking information of original meta search results is preserved in each class	ranking information of original meta search results is preserved in each class	ranking information of original meta search results is preserved in each class	conditional probability of the class given the document	Log Entropy	density of keywords
IndexUnit	word, bi-word and tri-word	word	word	word	word	word	word	none
IndexTech	morphology	morphology	morphology	morphology	morphology	maximal-extension indexing (original)	morphology	none
IndexStruc	inverted file	none	none	none	none	inverted file	inverted file	none
LabelType	predicted most frequetly term in the cluster	extracted feature terms based on their tfidf values	extracted feature terms based on their tfidf values	extracted feature terms based on their tfidf values	extracted feature terms based on their tfidf values	a machine-like id code followed by topical terms	none	for specialists' or 'for nonspecialists'
Resource	none	results of morphological analysis using Chasen	results of morphological analysis using Chasen	results of morphological analysis using Chasen	results of morphological analysis using Chasen	none	none	none

Details of each item at the first left column are explained in **Appendix**.

Figure 2. Summary of run result submission

Appendix: System Description Form

Each participating group was expected to submit a concise description of each classification run according to the following format:

- ⟨Subtask⟩ is fixed to 'D' in the Topical Classification Task.
- ⟨RunID⟩ identifies each run result in the manner of '⟨groupid⟩-⟨serialnumber⟩.cls,' *e.g.*, 'orgref-01.cls,' where the ⟨groupid⟩ indicates the group identification. The ⟨serialnumber⟩ indicates the serial number of the run.
- ⟨Method⟩ indicates whether the run is 'automatic' or 'interactive'. The 'automatic' and the 'interactive' indicate runs without any human intervention during classification run, and all runs other than 'automatic,' respectively.
- ⟨ClsModel⟩ Techniques used to classify the documents, *e.g.*, document clustering based on contents of documents, clustering based on hyperlink connections, clustering based on term co-occurrence within title and/or heading tags, text categorization based on a known subject/classification system, etc. Further details are welcome.
- ⟨LinkInfo⟩ specifies whether or not link information was used for classification processing, *e.g.*, link information only, link and contents information, contents only, etc.
- ⟨QueryInfo⟩ specifies whether or not query terms, i.e., TITLE of the topic statement were used for classification processing, and how. If the topic statement was not used, please fill in 'none'.
- ⟨Exclusion⟩ specifies whether or not exclusive classification was used, *e.g.*, exclusive classification or non-exclusive classification.
- ⟨Hierarchy⟩ specifies whether or not hierarchical classification was used, *e.g.*, hierarchical classification or non-hierarchical classification. The details should be described as the ClsModel.
- ⟨Ranking⟩ specifies the techniques used for document ranking within each class, *e.g.*, tf, tf-idf, mutual information, document length, PageRank, etc.
- ⟨LabelType⟩ Type of labels assigned to each class and/or the labeling method, *e.g.*, a machine-like identification code, topical terms, typical page titles, short summary sentences, snippets, etc.
- ⟨Resource⟩ specifies the the external resources used for classification processing or document ranking within each class, other than the data provided by organizers, *e.g.*, a known subject/classification system, training data set, etc.

⟨SYSDESC⟩
⟨SUBTASK⟩Subtask⟨/SUBTASK⟩
⟨RUNID⟩RunIDs⟨/RUNID⟩
⟨METHOD⟩Method⟨/METHOD⟩
⟨CLSMODEL⟩ClsModel⟨/CLSMODEL⟩
⟨LINKINFO⟩LinkInfo⟨/LINKINFO⟩
⟨QUERYINFO⟩QueryInfo⟨/ QUERYINFO⟩
⟨EXCLUSION⟩Exclusion⟨/EXCLUSION⟩
⟨HIERARCHY⟩Hierarchy⟨/HIERARCHY⟩
⟨RANKING⟩Ranking⟨/RANKING⟩
⟨LABELTYPE⟩LabelType⟨/LABELTYPE⟩
⟨RESOURCE⟩Resource⟨/RESOURCE⟩
⟨PRIORITY⟩Priority⟨/PRIORITY⟩
⟨INDEXUNIT⟩IndexUnit⟨/INDEXUNIT⟩
⟨INIDEXTECH⟩IndexTech⟨/INIDEXTECH⟩
⟨INDEXSTRUC⟩IndexStruc⟨/INDEXSTRUC⟩
⟨RUNTIME⟩RunTime⟨/RUNTIME⟩
⟨NOTE⟩Note⟨/NOTE⟩
⟨/SYSDESC⟩

Figure 3. Format of System Description

- ⟨Priority⟩ specifies the priority rank to each of four RunIDs, *e.g.*, 1, 2, 3 or 4, or '⟨runid⟩:1, ⟨runid⟩:2, ···,' when designating more than one RunIDs at once.
- ⟨IndexUnit⟩ optionally specifies the unit of index, *e.g.*, character, bi-character, word, bi-word, phrase, the name of the HTML tags used, link structure, etc.
- ⟨IndexTech⟩ optionally specifies the techniques used to process index terms, *e.g.*, morphology, stemming, POS, etc.
- ⟨IndexStruc⟩ optionally specifies the index structure, *e.g.*, PAT, inverted file, signature file, etc.
- ⟨RunTime⟩ optionally specifies the averaged seconds consumed for classifying.
- ⟨Note⟩ optionally specifies any other additional information.

Each system description should be flanked by '⟨SYSDESC⟩' and '⟨/SYSDESC⟩,' as shown in **Figure 3**. Each participating group was encouraged to describe each item of ⟨ClsModel⟩, ⟨Ranking⟩, ⟨IndexTech⟩, ⟨LabelType⟩ and ⟨Resource⟩ in detail and concretely, not limited to the examples indicated above.