Original Article

Healthc Inform Res. 2012 March;18(1):18-28.
http://dx.doi.org/10.4258/hir.2012.18.1.18
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Improving the Performance of Text Categorization Models used for the Selection of High Quality Articles

Seunghee Kim, BS, Jinwook Choi, MD, PhD
Department of Biomedical Engineering, Seoul National University College of Medicine, Seoul, Korea

Objectives: Machine learning systems can considerably reduce the time and effort needed by experts to perform new systematic reviews (SRs). This study investigates categorization models, which are trained on a combination of included and commonly excluded articles, which can improve performance by identifying high quality articles for new procedures or drug SRs. **Methods:** Test collections were built using the annotated reference files from 19 procedure and 15 drug systematic reviews. The classification models, using a support vector machine, were trained by the combined even data of other topics, excepting the desired topic. This approach was compared to the combination of included and commonly excluded articles with the combination of included and excluded articles. Accuracy was used for the measure of comparison. **Results:** On average, the performance was improved by about 15% in the procedure topics and 11% in the drug topics when the classification models trained on the combination of articles included and commonly excluded, were used. The system using the combination of included and commonly excluded articles performed better than the combination of included and excluded articles in all of the procedure topics. **Conclusions:** Automatically rigorous article classification using machine learning can reduce the workload of experts when they perform systematic reviews when the topic-specific data are scarce. In particular, when the combination of included and commonly excluded articles is used, this system will be more effective.

Keywords: Classification, Artificial Intelligence, Evidence-Based Medicine, Review Literature as Topic, Comparative Study

Submitted: February 1, 2012
Revised: February 29, 2012
Accepted: March 6, 2012

Corresponding Author
Jinwook Choi, MD, PhD
Department of Biomedical Engineering, Seoul National University College of Medicine, 103 Daehak-ro, Jongno-gu, Seoul 110-799, Korea. Tel: +82-2-2072-3421, Fax: +82-2-745-7870, E-mail: jinchoi@snu.ac.kr

I. Introduction

Evidence based medicine (EBM) is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients [1]. EBM is an important development in clinical practice and scholarly research [2].

Systematic review (SR) plays a key role in EBM [3]. SR attempts to identify, appraise and synthesize all the empirical evidences that meet pre-specified eligibility criteria to answer a given question [4]. Creation of a new SR or updating of an existing one takes considerable time and effort. First, the review topic and key questions are defined, and then relevant studies are retrieved from a number of different databases, such as MEDLINE and EMBASE. Next, experts select retrieved abstracts which are most likely to meet the inclusion criteria (abstract triage step). Finally, they closely read selected articles in the prior step and classify articles as included and excluded ones by pre-specified eligibility criteria (full text triage step) [5].

The new Health Technology Assessment (nHTA) center in National Evidence-based Healthcare Collaborating Agency (NECA) assesses new medical technologies introduced into Korean healthcare markets whether these technologies are safe and effective in real clinical settings or not. They review all the evidences systematically to evaluate those technologies. To date, 126 evidence reports have been completed and published [6].

Using current methods, we have not been able to cover new issues and keep even half of its reviews up-to-date [7]. We need to reduce avoidable processes in the production of research evidence [8]. Advanced information technologies can be developed and implemented to support SR by reducing the labor required while capturing high-quality evidence [3].

When an SR is first created, no data specific to this topic is available for information technologies. Cohen et al. [5] proposed a method that creates a model by training on data from a combination of other SR topics when topic-specific data is small. They compared to three systems, a baseline system using only topic-specific training data, a non-topic system using only the non-topic data sampled from the other topics and a hybrid system combining topic-specific training data with data from other SR topics. As the amounts of topic-specific training data become more available, their system preferentially incorporates these data into the model, reducing the influence of data from other topics. On average, the hybrid system improved mean area under the receiver operating characteristic (ROC) curves (AUC) over the baseline system by 20%, when topic-specific training data were scarce. In addition, the system performed better than the non-topic system at all but the two smallest fractions of topic specific training data. However, with very sparse topic-specific training data, the performance of the non-topic system on individual topics is often better than the baseline system, and is, at times, better than that of the hybrid system.

In this article, we address how the creation of SRs can be made more efficiently with machine learning (ML) techniques when the topic-specific data are small. We propose a method that creates classification models by training on articles included and commonly excluded. Inclusion and exclusion articles in SRs are judged by eligibility criteria. Those criteria are consisted of two parts; one is common exclusion criteria and the other is topic-specific inclusion/exclusion criteria. Articles excluded by common exclusion criteria are not included in other SRs regardless of topics. However, articles included or excluded by topic-specific criteria can be included or excluded in some SRs according to the topics. We hypothesized that by using commonly excluded articles across all SRs, we can automatically classify articles with better accuracy than previous works when a new SR is created.

II. Methods

We presented our methods in three parts. In the first, we described the data set used to evaluate our system. Then, we showed the classifier system and training method. Finally, we described our evaluation process.

1. Data Collection

In this study, the procedure data corpus was based on SR inclusion/exclusion judgments by the expert reviewers of the nHTA center. The expert reviewers classified articles at the abstract and full text level whether they are rigorous or not. This process is described in greater detail in earlier studies [5,9].

The reviewers classified articles by inclusion/exclusion criteria. Each article was encoded as shown in Table 1. Among criteria, there were 4 common exclusion criteria (code 1-4) across all SRs, such as grey literature (i.e., conference paper), non-original articles (i.e., review article, editorial, letter, and opinion pieces), non-human (animals) articles, and preclinical studies.

Among 126 SRs of nHTA, we selected 19 procedure SRs having more than 10 inclusion articles. Table 2 shows that the 19 review topics with the number of articles included and excluded in each study. We separated common exclusion articles (Excluded_com set) excluded by the common exclusion criteria (code 1-4) from exclusion articles (Excluded set) excluded by all the exclusion criteria (code 1-5).

Also, we used publicly available drug SRs to confirm our

Table 1. Coded values for article triage decisions in procedure topics

Code	Meaning
0	Included at article level
1	Excluded due to grey literature
2	Excluded due to non-original articles
3	Excluded due to non-human articles
4	Excluded due to pre-clinical studies
5	Excluded due to topic-specific reasons

Table 2. Number of articles included and excluded across 19 procedure systematic review topics

Topics	Included	Excluded[a]	Excluded_com[b]
Auditory brainstem implant	14	156	46
Autologous noncultured epidermal cellular transplantation	18	126	23
Continuous intraarticular pain control	22	742	38
Endoscopic cryotherapy of lung tumors	14	334	172
Glaucoma aqueous tube insertion	10	500	102
Hand transplantation	10	227	113
Holmium laser treatment of benign prostatic hyperplasia	34	155	93
Impedance controlled endometrial ablation	11	55	22
Intrastromal corneal ring surgery for keratoconus	31	140	26
Magnetic navigation assisted catheter technique	14	365	86
Radiofrequency ablation of primary and secondary lung malignancy	18	506	192
Small bowel transplantation	27	911	184
Somatic nerves stimulation	12	378	42
Surgical ablation of atrial fibrillation	13	185	66
Therapeutic temperature management with endovascular catheters	16	293	62
Therapeutic use of autologous bone marrow cells in peripheral arterial disease	28	249	143
Transanal endoscopic microsurgery	10	246	43
Transarterial radioembolization	32	473	156
Trigeminal nerve stimulation	11	730	50
Totals	345	6,771	1,659

[a]Exclusion articles excluded by all the exclusion criteria. [b]Exclusion articles excluded by the common exclusion criteria.

Table 3. Number of articles included and excluded across 15 drug systematic review topics

Topics	Included	Excluded[a]	Excluded_com[b]
ACEInhibitors	41	2,503	-
ADHD	20	831	1
Antihistamines	16	294	1
AtypicalAntipsychotics	146	974	11
BetaBlockers	42	2,030	104
CalciumChannelBlockers	100	1,118	25
Estrogens	80	288	-
NSAIDs	41	352	7
Opiods	15	1,900	-
OralHypoglycemics	136	367	-
ProtonPumpInhibitors	51	1,282	-
SkeletalMuscleRelaxants	9	1,634	-
Statins	85	3,380	-
Triptans	24	647	-
UrinaryIncontinence	40	287	21
Totals	846	17,887	170

ACE: angiotensin converting enzyme, ADHD: attention deficit hyperactivity disorder, NSAIDs: nonsteroidal antiinflammatory drugs.

[a]Excluded articles by all the exclusion criteria. [b]Articles excluded by the common exclusion criteria.

Table 4. Standardized coded values for article triage decisions in drug topics

Code	Meaning
I	Included at abstract or article level
E	Nonspecifically excluded
1	Excluded due to foreign language
2	Excluded due to wrong outcome
3	Excluded due to wrong drug
4	Excluded due to wrong population
5	Excluded due to wrong publication type
6	Excluded due to wrong study design
7	Excluded due to wrong study duration
8	Excluded due to background article
9	Excluded due to only abstract being available

method [10]. Tables 3 and 4 give information about the 15 drug topics and the inclusion/exclusion criteria [9]. We selected code 8 and 9 as common exclusion criteria across all drug SRs. Because, we thought background articles (code 8) might be non-original articles (i.e., review article, editorial, letter, and opinion pieces) and only abstract being available (code 9) might be grey literature (i.e., conference paper). We also separated common exclusion articles (code 8-9) from

http://dx.doi.org/10.4258/hir.2012.18.1.18

exclusion articles (code E-9). Looking at Table 3, the number of exclusion articles excluded by code 8 and 9 were small because most articles excluded by code E.

As can be seen in Tables 2 and 3, small number of articles was satisfied the inclusion criteria in most SRs. With small number of inclusion articles, it was not enough to the future predict. This approach may result in a model very biased towards negative (*Excluded*, as opposed to *Included*) prediction [5].

To prevent the biased prediction, we devised 'even' sets with the same number of inclusion and exclusion articles. We needed two even sets in one topic, because we divided exclusion articles into Excluded set and Excluded_com set. One was derived from Included and Excluded set (procedure/drug with Excluded even set) and the other was derived from Included and Excluded_com set (procedure/drug with Excluded_com even set). To make even sets, we randomly selected the same number of exclusion articles from Excluded set as inclusion articles if Excluded set had more articles than Included set. However, if Included set had more articles than Excluded set, we randomly selected the same number of inclusion articles from Included set as exclusion articles. For example, to make *Intrastromal Corneal Ring Surgery for Keratoconus* with Excluded even set, we randomly selected 31 exclusion articles from Excluded set with 140 articles, because Included set had 31 inclusion articles. This process yielded a total of 62 articles (31 exclusion and inclusion articles) as *Intrastromal Corneal Ring Surgery for Keratoconus* with Excluded even set. Also, to make *Intrastromal Corneal Ring Surgery for Keratoconus* with Excluded_com even set, we randomly selected 26 inclusion articles from Included set with 31 articles, because the Excluded_com set had 26 common exclusion articles. *Intrastromal Corneal Ring Surgery for Keratoconus* with Excluded_com even set had a total of 52 articles (26 exclusion and inclusion articles).

In the drug sets, we selected four topics (*AtypicalAntipsychotics, BetaBlockers, CalciumChannelBlockers, UrinaryIncontinence*) having more than 10 common exclusion articles, because some topics have very small number of articles in the Excluded_com sets. We also made even drug sets of four topics using the same method.

2. Classifier System

To determine the contribution of various feature types to the

Table 5. Number of training and test data across 19 procedure systematic review topics

Topics	With Excluded		With Excluded_com	
	Train	Test	Train	Test
Auditory brainstem implant	662	28	652	28
Autologous noncultured epidermal cellular transplantation	654	36	644	36
Continuous intraarticular pain control	646	44	636	44
Endoscopic cryotherapy of lung tumors	662	28	652	28
Glaucoma aqueous tube insertion	670	20	660	20
Hand transplantation	670	20	660	20
Holmium laser treatment of benign prostatic hyperplasia	622	68	612	68
Impedance controlled endometrial ablation	668	22	658	22
Intrastromal corneal ring surgery for keratoconus	628	62	628	52
Magnetic navigation assisted catheter technique	662	28	652	28
Radiofrequency ablation of primary and secondary lung malignancy	654	36	644	36
Small bowel transplantation	636	54	626	54
Somatic nerves stimulation	666	24	656	24
Surgical ablation of atrial fibrillation	664	26	654	26
Therapeutic temperature management with endovascular catheters	658	32	648	32
Therapeutic use of autologous bone marrow cells in peripheral arterial disease	634	56	624	56
Transanal endoscopic microsurgery	670	20	660	20
Transarterial radioembolization	626	64	616	64
Trigeminal nerve stimulation	668	22	658	22
Totals	12,420	690	12,240	680

classification task, we used four basic feature types as below: 1) words in the titles and abstracts of a MEDLINE citation; 2) Medical Subject Headings (MeSH) indexing terms from a MEDLINE citation; 3) publication types assigned manually by the National Library of Medicine (NLM) indexers.

The titles and abstracts were parsed into tokens. MeSH indexing terms and publication types were encoded these as phrases. Individual words of the titles and abstracts were further processed by removal of stop words such as 'the', 'an', and 'other' that are not likely to add semantic value to the classification [11]. The words were also stemmed by the Porter stemming algorithm, which reduced words to their roots [12].

As titles and abstracts were narrative text, the frequency-based representation worked better for them. On the other hand, since MeSH indexing terms and publication types did not occur in an article more than once, the binary representation method might be more suitable for the feature types [3]. Therefore, we represented the titles and abstracts by word frequencies and the MeSH indexing terms and publication types as binary.

To compare the various feature combinations for the classification tasks, we combined 4 features into 6 categories given below: 1) titles + abstracts (TA); 2) titles + abstracts + MeSH (TAM); 3) titles + abstracts + publication types (TAP); 4) titles + abstracts + MeSH + publication types (TAMP); 5) abstracts + MeSH + publication types (AMP); 6) MeSH + publication types (MP).

The ML system presented here was motivated by interesting results observed in earlier studies on finding the best evidence for SRs [2,13-15]. They noticed that using the support vector machine (SVM), rather than other MLs, led to improved classification performance. In the present work, our basic ML system was the SVM^{light} [16] implementation of the SVM algorithm, with a linear kernel and default settings [17].

The even set of each topic had small number of inclusion/exclusion articles. However, the accuracy of prediction systems based on a small number of sampled training data was unstable [18]. To solve this problem, we made training set combining even data of other topics except own topic about 4 collections (procedure/drug with Excluded set, procedure/drug with Excluded_com set). Because no data specific to new SR topic is available for information technologies, we did not include own topic data in training set. For example, to make *Auditory Brainstem Implant* training set, we combined even data of other 18 topics except the topic. Tables 5 and 6 show the number of training and test data across procedure/drug SR topics.

Table 6. Number of training and test data across 15 drug systematic review topics

Topics	With Excluded		With Excluded_com	
	Train	Test	Train	Test
ACEInhibitors	1,610	82	-	-
ADHD	1,652	40	-	-
Antihistamines	1,660	32	176	22
AtypicalAntipsychotics	1,400	292	-	-
BetaBlockers	1,608	84	114	84
CalciumChannelBlockers	1,492	200	148	50
Estrogens	1,532	160	-	-
NSAIDs	1,610	82	-	-
Opiods	1,662	30	-	-
OralHypoglycemics	1,420	272	-	-
ProtonPumpInhibitors	1,590	102	-	-
SkeletalMuscleRelaxants	1,674	18	-	-
Statins	1,522	170	-	-
Triptans	1,644	48	-	-
UrinaryIncontinence	1,612	80	156	42
Totals	23,688	1,692	594	198

ACE: angiotensin converting enzyme, ADHD: attention deficit hyperactivity disorder, NSAIDs: nonsteroidal antiinflammatory drugs.

3. Evaluation

We evaluated how well our categorization models which are trained on combination of included and commonly excluded articles perform on identifying rigorous articles for new procedure or drug SRs. In order to do that, first, we compared the classification accuracies using the various feature combinations in the procedure with Exclude set. Then, we compared the classification accuracies in the procedure/drug with Exclude set and the procedure/drug with Exclude_com set using the feature combination which shows the best classification accuracy in the procedure with Exclude set.

All collections were tested in the same processes. In the first step, we made 3 even sets in a topic; one is training set, others are test sets. We made two test sets, because randomly selected test data might affect performance results. In the second step, we combined training data of remaining topics except own topic. Finally, we built a general classification models by training on combined data of a given topic and classified 2 test sets of the topic. The accuracy was calculated for each constructed model, and all the computed results were averaged 2 test sets to give a final performance estimate. A representation of the overall process is shown in Figure 1.

We applied one-way ANOVA to compare classification accuracies of various feature combinations and t-test for results comparison of 4 collections. These statistical analyses used SPSS ver. 19 (SPSS Inc., New York, NY, USA).

III. Results

We presented the classification results of various feature combinations in the procedure with Exclude set and the accuracies in 4 collections using the best performance feature combination.

Table 7 shows the classification results of various feature combinations in the procedure with Exclude set. We found no statistical significance of the difference among them ($p > 0.05$). However, the MP showed the best accuracy, and was significantly better than the TAM ($p < 0.05$). With this result, we chose the MP as the best performance feature combination. Among topics, *Therapeutic Temperature Management with Endovascular Catheters* achieved the best average accuracies in 3 feature combinations (TAM, TAMP, MP) and *Small Bowel Transplantation* in others (TA, TAP, AMP).

Table 8 presents the results of procedure topics using the MP which is the best performance feature combination. We found that the overall mean percentage of accuracy in the procedure with Excluded_com set (88.32%) was significantly higher than that of the procedure with Excluded set (75.38%, $p < 0.05$). Also, all of topics in the procedure with Excluded_com set showed high or the same accuracies compared with those in the procedure with Exclude set.

Drug results are shown in Table 9. The overall mean percentage of accuracies in the drug with Excluded_com set was better than those of the drug with Excluded set. However, there was no statistically significant difference between them

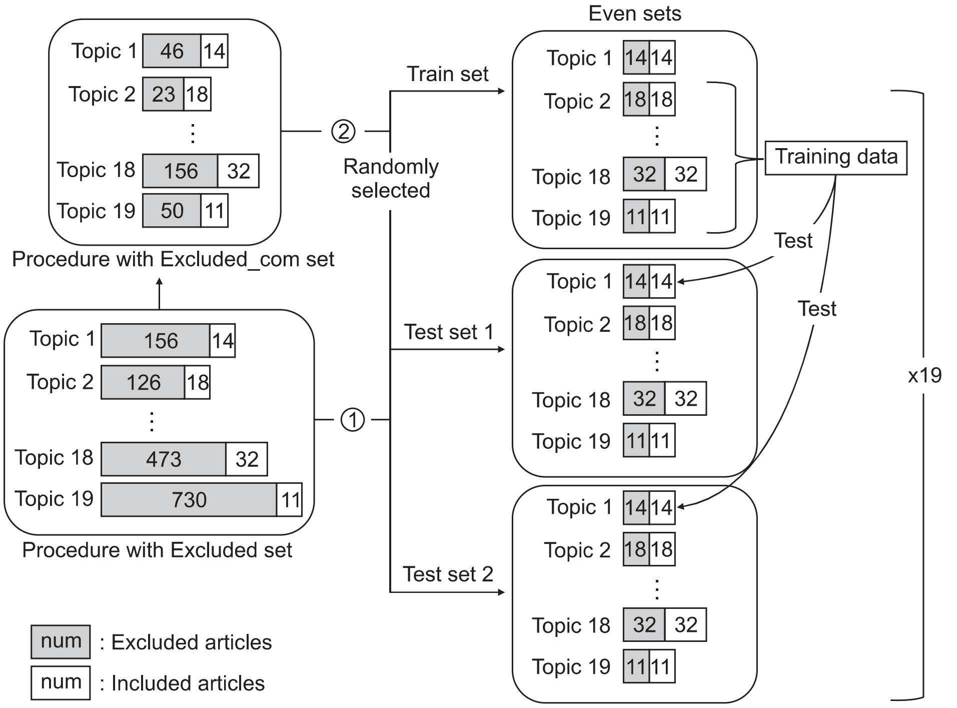

Figure 1. Evaluation processes of one topic.

Table 7. Mean percentage of various feature combinations accuracies in the procedure with Exclude set

Topics	TA	TAM	TAP	TAMP	AMP	MP
Auditory brainstem implant	66.08	67.86	75.00	69.65	66.07	73.22
Test set 1	64.29	71.43	71.43	71.43	60.71	78.57
Test set 2	67.86	64.29	78.57	67.86	71.43	67.86
Autologous noncultured epidermal cellular transplantation	61.11	63.89	66.67	68.06	62.50	69.45
Test set 1	61.11	66.67	66.67	69.44	63.89	66.67
Test set 2	61.11	61.11	66.67	66.67	61.11	72.22
Continuous intraarticular pain control	73.87	55.69	60.23	69.32	73.87	64.77
Test set 1	68.18	56.82	63.64	70.45	72.73	68.18
Test set 2	79.55	54.55	56.82	68.18	75.00	61.36
Endoscopic cryotherapy of lung tumors	62.50	60.72	62.50	55.36	60.72	80.36
Test set 1	64.29	64.29	60.71	57.14	64.29	78.57
Test set 2	60.71	57.14	64.29	53.57	57.14	82.14
Glaucoma aqueous tube insertion	60.00	67.50	65.00	65.00	70.00	72.50
Test set 1	55.00	70.00	65.00	70.00	75.00	70.00
Test set 2	65.00	65.00	65.00	60.00	65.00	75.00
Hand transplantation	65.00	62.50	62.50	72.50	62.50	50.00
Test set 1	70.00	60.00	65.00	75.00	65.00	50.00
Test set 2	60.00	65.00	60.00	70.00	60.00	50.00
Holmium laser treatment of benign prostatic hyperplasia	74.27	70.59	73.53	74.27	73.53	77.94
Test set 1	76.47	70.59	73.53	73.53	75.00	79.41
Test set 2	72.06	70.59	73.53	75.00	72.06	76.47
Impedance controlled endometrial ablation	63.64	63.64	70.46	63.64	63.64	52.28
Test set 1	63.64	63.64	68.18	63.64	63.64	54.55
Test set 2	63.64	63.64	72.73	63.64	63.64	50.00
Intrastromal corneal ring surgery for keratoconus	77.42	83.37	79.04	73.39	71.77	80.65
Test set 1	75.81	82.86	75.81	72.58	69.35	77.42
Test set 2	79.03	83.87	82.26	74.19	74.19	83.87
Magnetic navigation assisted catheter technique	60.71	55.36	64.29	64.29	55.36	82.14
Test set 1	60.71	57.14	64.29	64.29	53.57	82.14
Test set 2	60.71	53.57	64.29	64.29	57.14	82.14
Radiofrequency ablation of primary and secondary lung malignancy	72.22	56.95	68.06	68.06	65.28	81.95
Test set 1	72.22	58.33	75.00	66.67	63.89	86.11
Test set 2	72.22	55.56	61.11	69.44	66.67	77.78
Small bowel transplantation	83.33	76.86	84.26	77.78	83.33	75.93
Test set 1	83.33	75.93	81.48	75.93	83.33	79.63
Test set 2	83.33	77.78	87.04	79.63	83.33	72.22
Somatic nerves stimulation	70.83	58.34	66.67	62.50	77.09	68.75
Test set 1	70.83	62.50	62.50	62.50	75.00	70.83
Test set 2	70.83	54.17	70.83	62.50	79.17	66.67
Surgical ablation of atrial fibrillation	80.77	80.77	80.77	65.38	78.85	86.54
Test set 1	80.77	80.77	80.77	65.38	80.77	88.46
Test set 2	80.77	80.77	80.77	65.38	76.92	84.62
Therapeutic temperature management with endovascular catheters	67.19	84.38	64.07	78.13	73.44	92.19
Test set 1	65.63	84.38	62.50	78.13	68.75	93.75
Test set 2	68.75	84.38	65.63	78.13	78.13	90.63
Therapeutic use of autologous bone marrow cells in peripheral arterial disease	72.32	68.75	68.75	73.22	79.47	85.72
Test set 1	71.43	67.86	69.64	75.00	76.79	87.50
Test set 2	73.21	69.64	67.86	71.43	82.14	83.93
Transanal endoscopic microsurgery	57.50	57.50	62.50	65.00	62.50	82.50
Test set 1	60.00	60.00	75.00	70.00	70.00	80.00
Test set 2	55.00	55.00	50.00	60.00	55.00	85.00
Transarterial radioembolization	70.32	76.57	78.91	73.44	74.22	75.79
Test set 1	71.88	78.13	76.56	71.88	75.00	73.44
Test set 2	68.75	75.00	81.25	75.00	73.44	78.13
Trigeminal nerve stimulation	75.00	72.73	75.00	70.46	79.55	79.55
Test set 1	77.27	72.73	72.73	72.73	81.82	81.82
Test set 2	72.73	72.73	77.27	68.18	77.27	77.27
Mean	69.16	67.58	69.91	68.92	70.19	75.38

TA: titles + abstracts, TAM: titles + abstracts + MeSH, TAP: titles + abstracts + publication types, TAMP: titles + abstracts + MeSH + publication types, AMP: abstracts + MeSH + publication types, MP: MeSH + publication types.

Table 8. Mean percentage of accuracies in the procedure two sets using the MP

Topics	With Excluded	With Excluded_com
Auditory brainstem implant	73.22	73.22
Test set 1	78.57	78.57
Test set 2	67.86	67.86
Autologous noncultured epidermal cellular transplantation	69.45	88.89
Test set 1	66.67	88.89
Test set 2	72.22	88.89
Continuous intraarticular pain control	64.77	94.32
Test set 1	68.18	95.45
Test set 2	61.36	93.18
Endoscopic cryotherapy of lung tumors	80.36	83.93
Test set 1	78.57	82.14
Test set 2	82.14	85.71
Glaucoma aqueous tube insertion	72.50	92.50
Test set 1	70.00	95.00
Test set 2	75.00	90.00
Hand transplantation	50.00	60.00
Test set 1	50.00	60.00
Test set 2	50.00	60.00
Holmium laser treatment of benign prostatic hyperplasia	77.94	92.65
Test set 1	79.41	92.65
Test set 2	76.47	92.65
Impedance controlled endometrial ablation	52.28	79.55
Test set 1	54.55	77.27
Test set 2	50.00	81.82
Intrastromal corneal ring surgery for keratoconus	80.65	91.35
Test set 1	77.42	92.31
Test set 2	83.87	90.38
Magnetic navigation assisted catheter technique	82.14	92.86
Test set 1	82.14	92.86
Test set 2	82.14	92.86
Radiofrequency ablation of primary and secondary lung malignancy	81.95	91.67
Test set 1	86.11	91.67
Test set 2	77.78	91.67
Small bowel transplantation	75.93	89.82
Test set 1	79.63	90.74
Test set 2	72.22	88.89
Somatic nerves stimulation	68.75	95.83
Test set 1	70.83	95.83
Test set 2	66.67	95.83
Surgical ablation of atrial fibrillation	86.54	96.16
Test set 1	88.46	92.31
Test set 2	84.62	100.00
Therapeutic temperature management with endovascular catheters	92.19	95.32
Test set 1	93.75	93.75
Test set 2	90.63	96.88
Therapeutic use of autologous bone marrow cells in peripheral arterial disease	85.72	90.18
Test set 1	87.50	91.07
Test set 2	83.93	89.29
Transanal endoscopic microsurgery	82.50	95.00
Test set 1	80.00	90.00
Test set 2	85.00	100.00
Transarterial radioembolization	75.79	92.97
Test set 1	73.44	90.63
Test set 2	78.13	95.31
Trigeminal nerve stimulation	79.55	81.82
Test set 1	81.82	81.82
Test set 2	77.27	81.82
Mean	75.38	88.32

MP: MeSH + publication types.

Table 9. Mean percentage of accuracies in the drug two sets using the MP

Topics	With Excluded	With Excluded_com
AtypicalAntipsychotics	62.84	84.09
Test set 1	64.04	81.82
Test set 2	61.64	86.36
BetaBlockers	71.43	89.89
Test set 1	71.43	90.48
Test set 2	71.43	89.29
CalciumChannelBlockers	68.25	69.00
Test set 1	67.50	76.00
Test set 2	69.00	62.00
UrinaryIncontinence	70.63	65.48
Test set 1	71.25	64.29
Test set 2	70.00	66.67
Mean	68.29	77.12

MP: MeSH + publication types.

($p > 0.5$). All of topics in the drug with Excluded_com set showed higher accuracies than those in the drug with Excluded set, except *Urinary Incontinence*.

IV. Discussion

Our results showed that categorization models which are trained on combination of included and commonly excluded articles can improve articles triage performance to create SR. To improve the classification performance, we made even sets, combined even data of remaining topics except own topic, and used articles included and commonly excluded.

We compared classification accuracies of various feature combinations in the procedure with Exclude set. MP showed the best classification performance (75.38%), and the next was AMP (70.19%). TAM reported the worst classification performance (67.58%), and its overall mean percentage of accuracy was significantly lower than MP ($p < 0.05$). The metadata features like MeSH and publication types were benefiting the classification task, and the combination of all features (TAMP) was less successful than expected. Classification performances of feature combinations including title (TA, TAM, TAP, TAMP) were lower than other feature combinations (AMP, MP) without title. The feature combination showing the best performance (MP) was similar to that in the held-out test set to recognize methodologically rigorous studies of Kilicoglu et al. [15].

Using the best performance feature combination (MP), we tested our hypothesis that general categorization model, which is trained on the combination of included and commonly excluded articles, could classify methodologically rigorous articles with better accuracy than that which is trained on the simply combination of all articles. Overall mean accuracy of the procedure with Excluded_com set (88.32%) was significantly better than that of the procedure with Excluded set (75.38%) ($p < 0.05$). In all topics, accuracies of the procedure with Excluded_com set were higher than those in the procedure with Excluded set. Overall mean accuracy was 68.29% in the drug with Excluded set, and 77.12% in the drug with Excluded_com set. Except *UrinaryIncontinence*, accuracies of the drug with Excluded_com set were higher than those in the drug with Excluded set.

We analyzed articles encoded as 8 and 9 in 4 drug topics (*AtypicalAntipsychotics, BetaBlockers, CalciumChannelBlockers, UrinaryIncontinence*). Unlike our thought about code 8 and 9 in drug topics, publication types of articles encoded as 8 and 9 were various, such as Review, Comparative Study, and Clinical Trial. We classified those articles into three groups (articles excluded by our common exclusion criteria, topic-specific reasons, and other reasons). In *AtypicalAntipsychotics*, 45.45% of articles were excluded by our common exclusion criteria, 36.36% by topic-specific reasons, and 18.18% by other reasons. In *BetaBlocker*, 75.00% of articles were excluded by our common exclusion criteria, 19.23% by topic-specific reasons, and 5.77% by other reasons. In *CalciumChannelBlockers*, 48.00% of articles were excluded by our common exclusion criteria, 44.00% by topic-specific reasons, and 8.00% by other reasons. In *UrinaryIncontinence*, 23.81% of articles were excluded by our common exclusion criteria, 66.67% by topic-specific reasons, and 9.52% by other reasons. Three topics (*AtypicalAntipsychotics, BetaBlockers, CalciumChannelBlockers*) had more articles excluded by our common exclusion criteria than articles excluded by topic-specific and other reasons. However, *UrinaryIncontinence* had more articles excluded by topic-specific reasons. Classification accuracies of three topics (*AtypicalAntipsychotics, BetaBlockers, CalciumChannelBlockers*) having the large portion of articles excluded by our common exclusion criteria were improved in Excluded_com set. However, classification accuracy of *UrinaryIncontinence* having small portion of articles excluded by our common exclusion criteria was decreased in Excluded_com set.

There are several limitations to our evaluation. We randomly selected one training set and two test sets in a topic. Randomly selected one training data of a topic may not affect overall performance because we combined training data across topics except own topic. However, some randomly selected test data of a topic may poorly represent the topic and adversely affect performance. For example, *Hand Transplantation*, classification performance using MP was 50.00%. We thought test data of MP in *Hand Transplantation* might be poorly selected.

Our sample sizes are small. Although the data corpus includes 19 topics and expert judgments, overall articles are about 7,200. We used the data generated by a single SR-producing organization. It is also our limitation even if the nHTA uses the most rigorous processes to maximize quality and consistency. We tried to confirm our method using drug SRs generated by Drug Evidence Review Project (DERP) [5], but we could not evaluate our method with those articles properly. Because, in drug SRs, most articles were classified E (nonspecifically excluded) and all of articles with code 8 and 9 were not commonly excluded articles.

In conclusion, we have presented and evaluated a robust and effective method for improving the classification performance on articles for SRs. On average, performances were improved by about 15% in procedure topics and 11% in drug topics when categorization models, which are trained on combination of articles included and commonly excluded, were used. To the best of our knowledge, this is the first work in classification of scientifically rigorous studies using articles included and commonly excluded across all topics. Future work will focus on other classification features and classification algorithms to improve categorization performance.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education, Science and Technology (MEST) in Korea (No. 2011-0018259).

References

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312:71-2.
2. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. J Am Med Inform Assoc 2005;12:207-16.
3. Matwin S, Kouznetsov A, Inkpen D, Frunza O, O'Blenis P. A new algorithm for reducing the workload of experts in performing systematic reviews. J Am Med Inform Assoc 2010;17:446-53.
4. The Cochrane Library. About Cochrane systematic reviews and protocols [Internet]. West Sussex, UK: John Wiley & Sons, Ltd.; c2012 [cited at 2012 Mar 13]. Available from: http://www.thecochranelibrary.com/view/0/AboutCochraneSystematicReviews.html.
5. Cohen AM, Ambert K, McDonagh M. Cross-topic learning for work prioritization in systematic review creation and update. J Am Med Inform Assoc 2009;16:690-704.
6. Committee for New Health Technology Assessment. nHTA. Seoul, Korea: Ministry of Health and Welfare; c2012 [cited at 2012 Mar 13]. Available from: http://neca.re.kr/nHTA/english/.
7. Koch G. No improvement - still less than half of the Cochrane reviews are up to date. In: 14th Cochrane Colloquium, 2006.
8. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet 2009;374:86-9.
9. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assoc 2006;13:206-19.
10. Cohen AM. Systematic drug class review gold standard data [Internet]. Portland (OR): Oregon Health & Science University; c2010 [cited at 2011 May 16]. Available from: http://davinci.ohsu.edu/~cohenaa/systematic-drug-class-review-data.html.
11. Onix text retrieval toolkit: API reference [Internet]. Provo (UT): Lextek International; c2000 [cited at 2011 May 21]. Available from: http://www.lextek.com/manuals/onix/stopwords1.html.
12. Porter MF. An algorithm for suffix stripping. Program 1980;14:130-7.
13. Cohen AM. Optimizing feature representation for automated systematic review work prioritization. AMIA Annu Symp Proc 2008;2008:121-5.
14. Joachims T. Text categorization with support vector machines: learning with many relevant features. In: Proceedings of the 10th European Conference on Machine Learning, 1998. p.137-42.
15. Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards automatic recognition of scientifically rigorous clinical research evidence. J Am Med Inform Assoc 2009;16:25-31.

16. Joachims T. Support vector machine: SVMlight [Internet]. Ithaca (NY): Cornell University; c2008 [cited at 2011 May 20]. Available from: http://svmlight.joachims.org/.
17. Joachims T. Making large-scale support vector machine learning practical. In: Scholkopf B, Burges CJ, Smola AJ, eds. Advances in kernel methods. Cambridge (MA): MIT Press; 1999. p.169-84.
18. Lee YH, Cheng TH, Lan CW, Wei CP, Hu PJ. Overcoming small-size training set problem in content-based recommendation: a collaboration-based training set expansion approach. In: Proceedings of the 11th International Conference on Electronic Commerce, 2009. p.99-106.

http://dx.doi.org/10.4258/hir.2012.18.1.18