Subdomain Entry Vocabulary Modules Evaluation

Technical Report

Vivien Petras
August 11, 2000

Abstract: Subdomain entry vocabulary modules represent a way to provide a more specialized retrieval vocabulary in a particular subject area. Several subdomain indexes have been derived for an analysis using the INSPEC database. The results show that subdomain indexes differ significantly from each other and from the general-purpose index they were derived from. The document pools that could be retrieved using the different subdomain entry vocabulary modules also differ greatly. If a word can be understood in more than one sense (polysemy), it is more likely to lead to different output from the individual subdomain indexes.
Evaluation of the prediction power of subdomain Entry Vocabulary Modules shows that more specific Entry Vocabulary Modules are more precise in predicting correct subject headings for given documents in a subject area.

Related papers and reports:
Michael K. Buckland, Aitao Chen, Michael Gebbie, Youngin Kim, & Barbara Norgard. Variation by Subdomain in Indexes to Knowledge Organization Systems.
http://www.sims.berkeley.edu/research/metadata/iskopaper.html
Youngin Kim. Evaluation of the Sensitivity of Subdomain in EVM dictionary approach Technical Report, 2000.
http://www.sims.berkeley.edu/research/metadata/papers/subdomain00.html
Youngin Kim. Evaluation of the performance of the EVM dictionaries. Technical Report, 2000.
http://www.sims.berkeley.edu/research/metadata/eval_desc.htm
Vivien Petras. Variation on Subdomain Indexes Technical Report, 2000.
http://www.sims.berkeley.edu/research/metadata/papers/subdvariation.html

I. Introduction

1. Subdomain Entry Vocabulary Modules

Subdomain Entry Vocabulary Modules (EVMs) are specialized indexes derived from a general-purpose index in order to represent a smaller and more qualified search vocabulary for knowledge systems in certain research areas. We refer to these indexes as entry vocabulary modules because they help a user finding appropriate search terms for formulating a query strategy for the knowledge system. Entry Vocabulary Indexes that cover a specific subdomain or research area embrace the specialized vocabulary of this subject and reflect the specialized language in their predictions for appropriate query terms.

Subdomain Entry Vocabulary Modules will understand the user's language and his information need and provide him with search terms (thesaurus terms, subject headings) appropriate for his subject area.
An Entry Vocabulary Module is created by forming a dictionary of associations between lexical items found in the titles, authors, and/or abstracts of existing records linked to the subject area in the knowledge system. A likelihood ratio statistic is used to measure association between these and to predict which of the metadata terms (i.e. classification numbers, subject headings, or thesaurus terms) best mirror the topic represented by the searcher's search vocabulary. This technique was developed under the name "Classification clustering" by Ray Larson for the Cheshire Information Retrieval system[1] and later further developed to incorporate natural language processing[2] for computing associations between noun phrases instead of only individual words. A more detailed account of how EVMs are created can be found in another place[3].

2. This report

This report will explain several experiments that were conducted to compare specialized subdomain indexes (EVMs) with a general index. If the subdomain indexes indeed provide more purposeful retrieval terms than a general index, then the subdomain EVM can be regarded as a very helpful tool in the retrieval process.
We compared subdomain indexes with the general index with regard to their variability in suggesting search terms and subsequent document pools that were researched with these search terms.
In a second series of experiments we tried to evaluate the prediction power of subdomain Entry Vocabulary Modules. We measured precision and recall values of subdomain EVMs for predicting correct subject headings for given bibliographic records of journal articles.

3. Source Data for building EVMs

As a source for building the Entry Vocabulary Modules we used the INSPEC database. INSPEC is an abstracting service covering over 4,000 scientific journals, conference proceedings, books, reports, and dissertations in the subject areas of Physics, Electrical and Electronic Engineering, Computers and Control, and Information Technology. We used the INSPEC dataset available from the University of California Digital Library in association with the Melvyl online catalog.
There are several strategies for defining a subject area (in order to build an EVM) imaginable. For building EVMs that would represent a general index we used randomly retrieved records from the INSPEC database. This would allow to generate EVMs that provide a general image of the vocabulary of the whole database.

---

[1] Larson, R. (1991): Classification Clustering, Probabilistic Information Retrieval and the Online Catalog. Library Quarterly, vol. 61, no. 2, p. 133-173
[2] Kim, Y. and Norgard, B. (1998): Adding Natural Language Processing Techniques to the Entry Vocabulary Module Building Process. Technical Report
http://www.sims.berkeley.edu/research/metadata/nlptech.html
[3] Plaunt, C. and Norgard, B. (August 1998): An Association Based Method for Automatic Indexing with a Controlled Vocabulary, JASIS, Vol. 49, no. 10, p. 888-902

For defining more specific subject areas we used two different strategies. The Science Citation Index Journal Citation Report cites a list of important journals for many different subject areas. We used this report as an authoritative resource for determining subject areas and important journals that cover these areas. In a second step we retrieved records from the INSPEC database that described articles authored in these journals and build a subdomain EVM with these records. This strategy was used in the first series of experiments (Variation of Subdomain Indexes).
The INSPEC database uses a classification system besides its subject headings to describe its bibliographic records. It is divided into four main sections: A Physics, B Electrical & Electronic Engineering, C Computers & Control, and D Information Technology, which are divided into further sub-categories indicated by a decimal number system. We used classification categories to determine subject areas within the INSPEC database. Subdomain EVMs were created by using records that would appear within the same classification category. This strategy was used in the second series of experiments (Evaluation of the prediction power of Subdomain Indexes).

For the first series of experiments (Variation of Subdomain Indexes), we created a "General" Entry Vocabulary Index based on a random sample of 152,646 INSPEC records. We also created three subdomain indexes:
- "Biotechnology", of records from journals listed in the Science Citation Index Journal Citation Report subject category "Biotechnology and Applied Microbiology"[4]
- "Information Science", using 9,549 records (retrieved in August 1998) from journals listed in the Science Citation Index Journal Citation Report subject category "Information Science and Library Science", and
- "Water", using 9,613 records (retrieved June 1999) from journals listed in the Science Citation Index Journal Citation Report subject category "Water Resources".
These EVMs are available for searching at
http://www.sims.berkeley.edu/research/oasis.html

The EVMs for the second series of experiments will be described in the respective sections.

II. Variation of Subdomain Indexes

1. Experiment I: How different are subdomain indexes from a general index?

We created random sets of sample words to test whether the subdomain indexes would suggest different thesaurus terms for the sample terms than the general index.
We created sample sets with 600 words that were taken randomly from the dictionaries of the four EVMs (General, Biotechnology, Information Science, and Water). The words were checked against WordNet 1.6, an online thesaurus that enumerates the different meanings (senses) of each word. A sample set was composed of 100 words with a single meaning, 100 words with two meanings, and 100 words with three, four, five, and six meanings.

[4] Unfortunately, we don't know how many records are in this EVM because the source data seems to be lost.

The sample from the general Index was then used as query to search against the general Index and the three subdomain indexes. In many cases one of the subdomain indexes did not contain a thesaurus term for the sample term. These sample terms were discarded. For the remaining 127 sample words, the number of different thesaurus terms (from index to index) was counted.
The difference was significant. In 70.8% of the cases (90 out of 127) the three subdomain indexes suggested a different thesaurus term than the general index. In 22.8% (29 out of 127) of the cases, two subdomain EVMs yielded different terms, and in 6.3% of the queries (8 out of 127) only one index had a different thesaurus term. For this sample set of words, in none of the cases all three subdomain indexes would yield the same thesaurus term than the general index.[5]

Experiment I (Subdomain EVMs vs. General Index)		
Sample terms	127	Out of 600
One EVM equal to general index	29	22.83%
Two EVMs equal to general index	8	6.30%
Three EVMs equal to general index	0	0.00%
All 4 different	90	70.87%

We repeated this experiment with three random EVMs that were twice as big as the subdomain EVMs (20,000 records) and compared their suggested thesaurus terms with those the General index would suggest. As predicted, the difference between random EVMs (that actually could be considered as smaller clones of a general index) and the general index was much less marked. From the 326 sample terms that remained after all the cases, where a subdomain EVM didn't find a thesaurus term, were discarded, 15.6% would lead to the same thesaurus term from the subdomain indexes as well as the general index. In 8.9% of the cases, two subdomain indexes would yield the same thesaurus term as the general index, and in 15.3% of the cases at least one subdomain EVM would propose the same thesaurus term.

Experiment I (Random EVMs vs. General Index)		
Sample terms	326	Out of 600
One EVM equal to general index	50	15.34%
Two EVMs equal to general index	29	8.90%
Three EVMs equal to general index	51	15.64%
All 4 different	196	60.12%

[5] It would be interesting to compare the number of unique subject headings each EVM has and also the overlap between EVMs regarding their subject headings. It wasn't possible for me to obtain the number of unique subject headings for the subdomain EVMs but the general EVM has 8,311 unique subject headings in its dictionary.
The number of unique subject headings and their overlap in the individual EVMs probably plays an important role in determining how many times the EVMs will suggest the same thesaurus terms for given sample terms, especially as the number of unique subject headings does not grow proportionally with the size of the EVMs and the probability for yielding the same thesaurus term increases. See later footnote for more details.

2. Experiment Ia: Index variability and index size

We compared how subdomain EVMs would differ from each other in suggesting thesaurus terms. Later, we compared the subdomain EVMs to randomly created EVMs (resembling the general index) to make sure that subdomain EVMs generate different results than general EVMs of the same size. For experiment Ia we submitted one sample of 600 words from the Information Science index, one sample of 600 words from the Biotechnololgy index, and one sample of 600 words from the Water index to all three subdomain EVMs and compared the thesaurus terms that were suggested. As in experiment I, we discarded all cases where one EVM wouldn't find a thesaurus term. We repeated this experiment with another sample set of terms (again 600 words from each subdomain index) and got surprisingly similar results.
In ca. 90% of the cases, all three subdomain EVMs would yield a different thesaurus term for a given sample term. In about 8% two subdomain EVMs would have the same thesaurus term, and in only about 1% of the cases would all three subdomain EVMs suggest the same thesaurus term.

Experiment Ia (Subdomain EVMs compared)		
Consolidated from 3 sample set with 600 words each, Set M		
Sample terms:	804	Out of 1800
Two EVMs equal	70	8.71%
Three EVMs equal:	10	1.24%
All 3 EVMs different:	724	90.05%

Experiment Ia (Subdomain EVMs compared)		
Consolidated from 3 sample set with 600 words each, Set V		
Sample terms:	840	Out of 1800
Two EVMs equal	67	7.98%
Three EVMs equal:	10	1.19%
All 3 EVMs different:	763	90.83%

These results can be compared with randomly created EVMs with the about the same size (number of records ~ 10,000) of the subdomain EVMs. As predicted, random EVMs are much more alike: the probability that all three EVMs would suggest the same thesaurus is much higher (ca. 16%). In only 65% would each random EVM suggest a different thesaurus term, and in 19% of the cases two EVMs would suggest the same thesaurus term.

Experiment Ia (Random EVMs with 10,000 records compared)		
Consolidated from 3 sample sets with 600 words each, Set M		
Sample terms:	1250	Out of 1800
Two EVMs equal	241	19.28%
Three EVMs equal:	201	16.08%
All 3 EVMs different:	808	64.64%

Experiment Ia (Random EVMs with 10,000 records compared)		
Consolidated from 3 sample sets with 600 words each, Set V		
Sample terms:	1242	Out of 1800
Two EVMs equal	244	19.65%
Three EVMs equal:	197	15.86%
All 3 EVMs different:	801	64.49%

Comparing randomly created EVMs with increasing size (number of records indexed for association dictionary), it became clear that the bigger an EVMs, the more it resembles the general index for the knowledge system. It is subsequently more likely for three EVMs to suggest the same thesaurus terms for a given sample term. For subdomain EVMs, however, we still expect more variability in suggesting thesaurus terms because they don't resemble the general index as much – even with a bigger size.

Experiment Ia (Random EVMs with 20,000 records compared)		
Consolidated from 3 sample sets with 600 words each, Set M		
Sample terms:	1416	Out of 1800
Two EVMs equal	266	18.79%
Three EVMs equal:	293	20.69%
All 3 EVMs different:	857	60.52%

Experiment Ia (Random EVMs with 80,000 records compared)		
Consolidated from 3 sample sets with 600 words each, Set M		
Sample terms:	1603	Out of 1800
Two EVMs equal	389	24.27%
Three EVMs equal:	523	32.63%
All 3 EVMs different:	691	43.11%

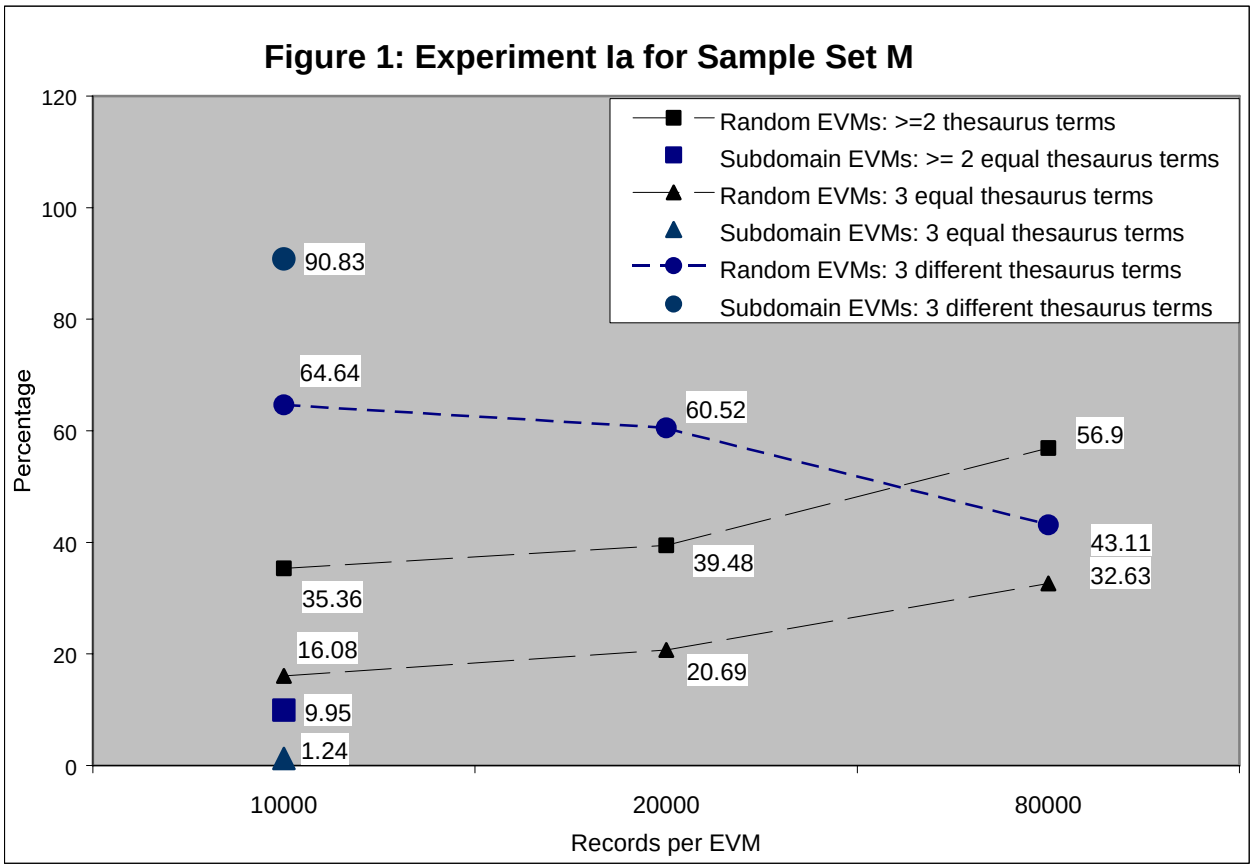
Experiment Ia (Random EVMs with 20,000 records compared)		
Consolidated from 3 sample sets with 600 words each, Set V		
Sample terms:	1391	Out of 1800
Two EVMs equal	283	20.35%
Three EVMs equal:	276	19.84%
All 3 EVMs different:	832	59.81%

Experiment Ia (Random EVMs with 80,000 records compared)		
Consolidated from 3 sample sets with 600 words each, Set V		
Sample terms:	1576	Out of 1800
Two EVMs equal	380	24.11%
Three EVMs equal:	530	33.63%
All 3 EVMs different:	666	42.26%

It is clear from these numbers that bigger EVMs reflect the characteristics of the general index much more and are therefore more similar to each other (in this case, the similarity leads to a higher probability of suggesting the same thesaurus terms). It is curious that in the experiment with the biggest random EVMs (80,000 records) the cases where three EVMs suggest the same thesaurus term are more often than the cases where only two EVMs suggest the same thesaurus term. This could be explained with the resemblance of the three EVMs and the higher likelihood to predict the same thesaurus term for a given sample term for all three EVMs than only two[6].

[6] This could also be explained with the number of unique subject headings and their overlap in bigger EVMs. Although the size of the EVMs doubled and quadrupled, the number of unique subject headings grew very slowly. For the 3 random EVMs with 10,000 records the average number of unique subject headings was 6,110 (6,116; 6,109; 6,104). For random EVMs with 20,000 records the average number of unique subject headings was 6,837 (6,830; 6,843; 6,838) and for random EVMs with 80,000 records the average number of unique subject headings was 7,689 (7,694; 7,683; 7,689).

For a graphical display of experiment Ia for sample set M see Figure 1.



3. Experiment II: Multiple meanings and index variability

In this experiment, we measured how much the polysemy of words would influence the variability of the subdomain indexes. Each sample word was searched against all three subdomain indexes, which resulted in a "variability" on a scale from 1 to 3 according to whether one, two, or three different thesaurus terms were suggested by the indexes. From the process of sampling we already knew that each word of the samples had a certain number of senses (from WordNet)[7].
Our hypothesis was that the more meanings a word has, the more likely it is that the three subdomain EVMs would suggest different thesaurus terms for a given sample term because it is more likely that the different subject areas use the word with a different meaning.

Two strategies were employed to calculate the relations between EVM variability and polysemy. The first method would consider the “not found” cases (where one or more EVMs didn’t find a thesaurus term) as similar to EVMs finding the same thesaurus terms

[7] Our samples from the subdomain indexes consist of 600 words with each 100 words with a single meaning (sense in WordNet), 100 words with 2 senses, 100 words with 3 senses, 100 words with 4 senses, 100 words with 5 senses, and 100 words with 6 senses.

and reduce the variability by 1 each time a “not found” case appeared. The second method would discard all “not found” cases and only keep cases where all three EVMs found a thesaurus term. The following table gives an overview of how the variability was calculated with these two methods.

MG method
including "not found" cases

EVM 1	EVM 2	EVM 3	Variability
t1	t2	t3	3
t1	t2	t1	2
t1	t2	N	2
t1	t1	t1	1
t1	t1	N	1
t1	N	N	1

VP method
discarding "not found" cases

EVM 1	EVM 2	EVM 3	Variability
t1	t2	t3	3
t1	t2	t1	2
t1	t2	N	not applicable
t1	t1	t1	1
t1	t1	N	not applicable
t1	N	N	not applicable

t1, t2, t3 = thesaurus terms
N = “not found” case

There are arguments for both strategies. One could argue that the VP method leaves out valuable information like whether only one or two EVMs didn’t find the equivalent thesaurus term for a given sample term. On the other hand, the MG method distorts the results towards the variabilities of one or two because the “not found” cases automatically reduce the variability by 1 so that a variability of three is much less likely (especially considering that over 50% of all sample term don’t yield a thesaurus term from one or the other EVM).

We compared again subdomain EVMs and randomly created EVMs with each other[8]. Like in experiments I and Ia we submitted sample sets of 600 words from each subdomain index (Info, Bio, Water) to all of the three subdomain EVMs. We then calculated the variability (how many equal thesaurus terms were suggested) for each sample term. The second step involved calculating the average number of senses for each variability constant.

We found that our hypothesis was confirmed for the MG method. The VP method found only a weak relation and didn’t confirm the hypothesis for our second sample set V. However, subdomain EVMs showed a greater deviation than the randomly created EVMs.
It might be necessary to statistically confirm the significance of the shown trends.

Figures 2-5 show the results for this experiment.

[8] The random EVMs have a size of about 20,000 records.

Figure 2: Experiment II, MG Method, M Sample

MG Method (including "not found" cases)	Results from Subdomain EVMs			Results from Random EVMs		
	Terms from Info sample*	Terms from Bio sample	Terms from Water sample	Terms from Info sample	Terms from Bio sample	Terms from Water sample
Variability	**Average number of senses**					
0**				2.53	2.58	2.62
1	3.03	2.56	2.86	3.52	3.46	3.49
2	3.39	3.13	3.29	3.45	2.7	3.55
3	4.15	4.09	4.15	3.64	3.84	3.56

* 600 sample terms
** A variability of 0 could occur when none of the EVMs would suggest a thesaurus term and the MG method reduced the variability by 3.

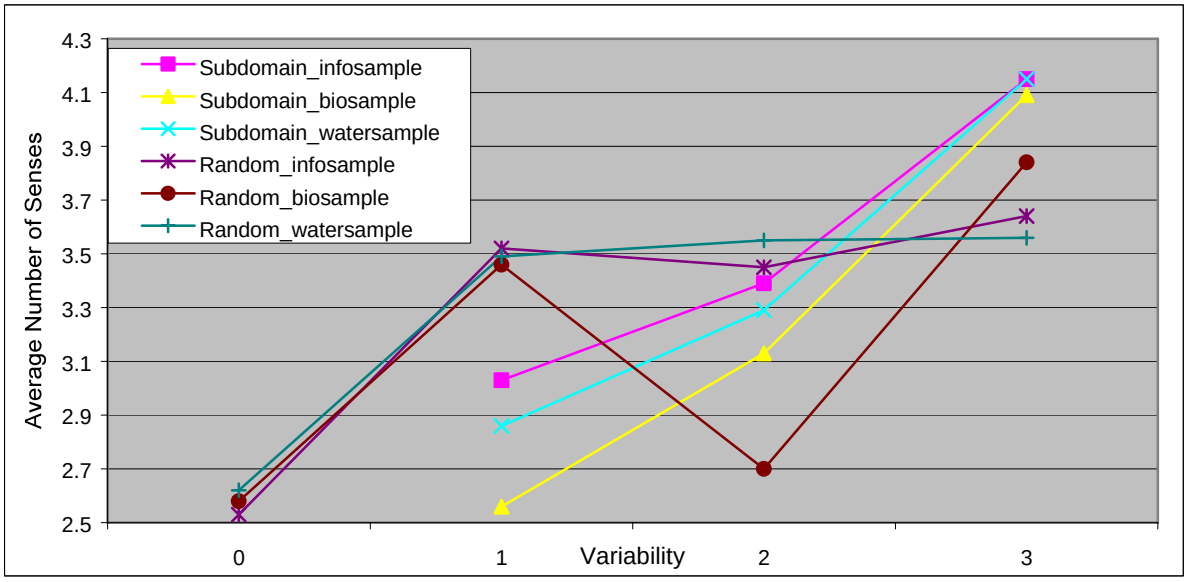


Figure 3: Experiment II, MG Method, V Sample

MG Method (including "not found" cases)	Results from Subdomain EVMs			Results from Random EVMs		
	Terms from Info sample	Terms from Bio sample	Terms from Water sample	Terms from Info sample	Terms from Bio sample	Terms from Water sample
Variability	**Average number of senses**					
0				2.38	2.23	3.13
1	2.78	2.25	3.13	3.42	3.33	3.44
2	3.58	3.37	3.43	3.61	3.72	3.65
3	3.96	4.13	3.82	3.64	4.06	3.51

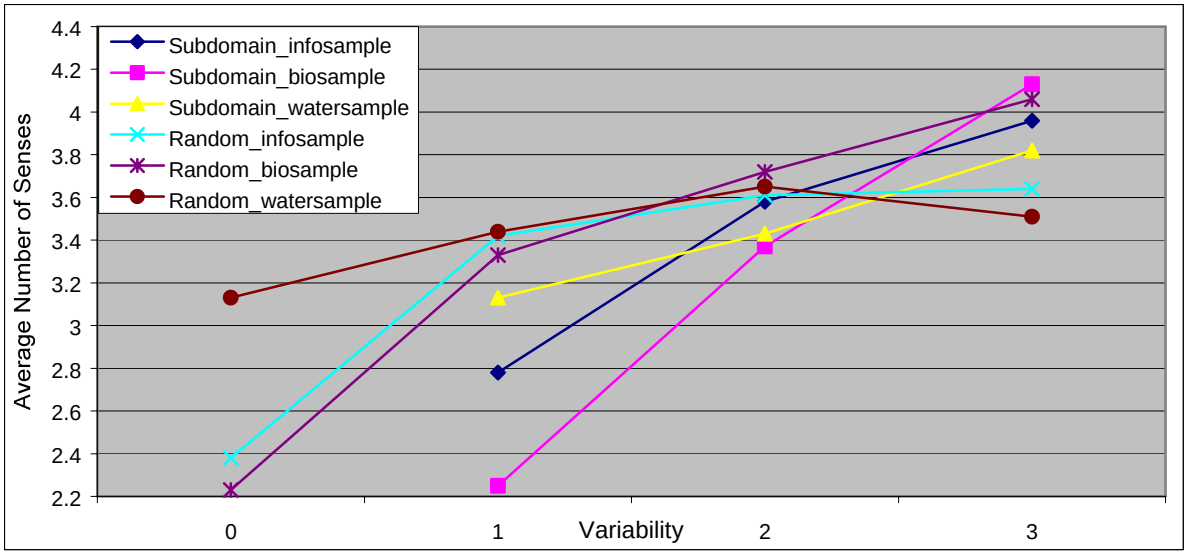


Figure 4: Experiment II, VP Method, M Sample

VP Method (discarding "not found" cases)	Results from Subdomain EVMs			Results from Random EVMs		
	Terms from Info sample*	Terms from Bio sample	Terms from Water sample	Terms from Info sample	Terms from Bio sample	Terms from Water sample
Variability	**Average number of senses**					
1	2.5	3.75	2.75	3.83	3.75	3.81
2	3.62	3.84	4	3.82	3.76	3.68
3	4.13	3.99	4.15	3.64	3.83	3.56

*Terms left after discarding “not found” cases:

- Subdomain EVMs: Info sample=191, Bio sample=377, Water sample=236
- Random EVMs: Info sample=456, Bio sample=454, Water sample=506

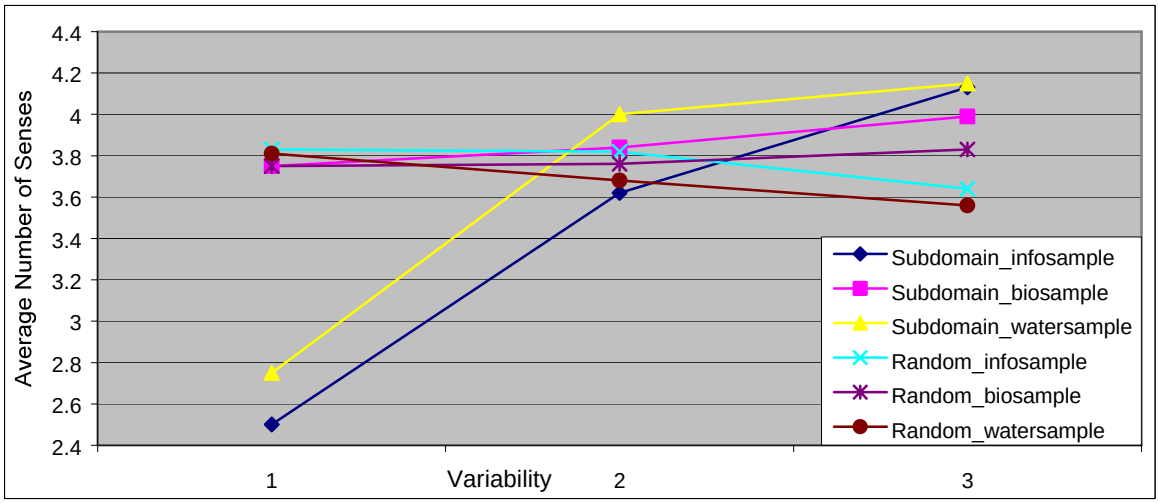
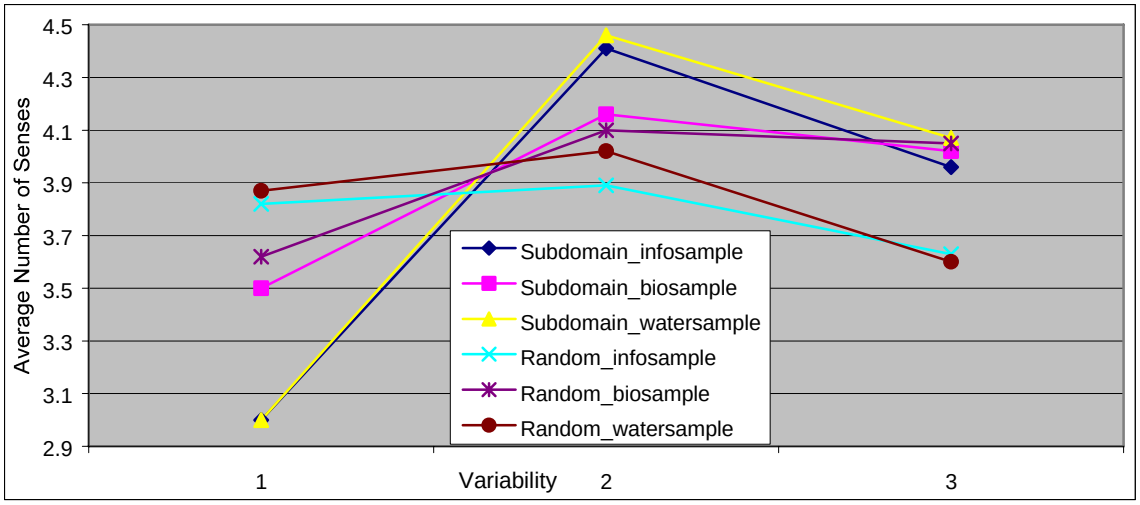


Figure 5: Experiment II, VP Method, V Sample

VP Method (discarding "not found" cases)	Results from Subdomain EVMs			Results from Random EVMs		
	Terms from Info sample*	Terms from Bio sample	Terms from Water sample	Terms from Info sample	Terms from Bio sample	Terms from Water sample
Variability	**Average number of senses**					
1	3	3.5	3	3.82	3.62	3.87
2	4.41	4.16	4.46	3.89	4.1	4.02
3	3.96	4.02	4.07	3.63	4.05	3.6

*Terms left after discarding "not found" cases:

- Subdomain EVMs: Info sample=233, Bio sample=383, Water sample=224
- Random EVMs: Info sample=473, Bio sample=432, Water sample=486



4. Experiment III: How different are the search results of subdomain EVM retrieval?

To further confirm the results from experiment I and Ia (where we analyzed how many common thesaurus terms are suggested by different subdomain Entry Vocabulary Modules) we questioned now how big the overlap in documents, which could actually be retrieved, is. We examined the document pools that could be retrieved using the suggested thesaurus terms from the three special subdomain entry vocabularies Biotechnology, Information Science, and Water.
Firstly, the sample terms (the same sets we also used in the previous experiments) were submitted to the EVMs to retain the top preferred thesaurus term. We discarded the cases where one EVM didn't find an equivalent thesaurus term for a given sample term. The suggested thesaurus terms were then submitted to the INSPEC database on Melvyl to retrieve the actual documents (containing the suggested thesaurus terms). By applying a Boolean query strategy we could find the documents that had more than one of the suggested thesaurus terms in common. For each sample term and its subsequent three

suggested thesaurus terms (one Biotechnology thesaurus term, one Information science thesaurus term, and one Water thesaurus term) we submitted the following 7 queries to INSPEC:

1.number of documents found with the thesaurus term from the Information science EVM
2.number of documents found with the thesaurus term from the Biotechnology EVM
3.number of documents found with the thesaurus term from the Water EVM
4.number of documents found with the thesaurus terms from the Information science AND Biotechnology EVMs (intersection)
5.number of documents found with the thesaurus terms from the Biotechnology AND Water EVMs (intersection)
6.number of documents found with the thesaurus terms from the Information science AND Water EVMs (intersection)
7.number of documents found with the thesaurus terms from the Information science AND Biotechnology AND Water EVMs (intersection).

In order to examine the impact of loose and rigid query strategies we applied two query strategies:

i) rigid query strategy (restrict the number of documents found)
requiring the occurrence of the sample term together with the suggested thesaurus term in the same document
e.g. sample term = galileo, suggested thesaurus term by the Information science EVM = reservation computer systems
query # 1 = FI KW galileo AND XSU reservation computer systems

ii) loose query strategy
requiring only the occurrence of the suggested thesaurus term in the controlled or free subject headings of the document.
e.g. sample term = galileo, suggested thesaurus term by the Information science EVM = reservation computer systems
query # 1 = FI SU reservation computer systems

The results were astounding. The overlap between documents resulting from queries from different subdomain EVM thesaurus terms is very small: for the rigid query strategy, 4.16% of the documents retrieved contained all three suggested thesaurus terms (from the three EVMs) and the sample term. Interestingly, for the loose query strategy the number was even smaller (1.07%). Only very few sample terms (1-6) per sample file actually account for the greatest part of this overlap (e.g. sample terms that lead to the same top index terms in all three EVMs and retrieve a lot of documents).
In general, queries requiring the Information science AND Biotechnology EVM thesaurus terms have more documents in common (22.17% for rigid, 5.73 for loose query strategy) than queries requiring the Biotechnology AND Water EVM thesaurus terms (18.90% for rigid, 4.53% for loose query strategy), which in turn have more documents in

common than those requiring the Information science AND Water EVM thesaurus terms (10.63% for rigid, 3.28% for loose query strategy).

Experiment III Rigid Query Mode M sample set*	Query 1 (info thesaurus Term)	Query 2 (bio thesaurus term)	Query 3 (water thesaurus term)	Query 4 (intersection info AND bio)	Query 5 (intersection bio AND water)	Query 6 (intersection info AND water)	Query 7 (intersection all thesaurus terms)
# of Documents from info sample	100320	138662	117145	41804	27167	10596	4516
# of Documents from bio sample	275632	358866	344455	159365	116536	74917	54127
# of Documents from water sample	112770	154501	140237	37244	57511	11658	9452
	488722	652029	601837	238413	201214	97171	68095
Sample terms: 804							
Average	608	811	749	297	250	121	85
sum				1122	1309	1236	2083
Overlap of documents with thesaurus terms from several EVMs				**26.42%**	**19.11%**	**9.78%**	**4.07%**

*Terms left after discarding "not found" cases:

- Subdomain EVMs: Info sample=191, Bio sample=377, Water sample=236

Experiment III Rigid Query Mode V sample set**	Query 1 (info thesaurus Term)	Query 2 (bio thesaurus term)	Query 3 (water thesaurus term)	Query 4 (intersection info AND bio)	Query 5 (intersection bio AND water)	Query 6 (intersection info AND water)	Query 7 (intersection all thesaurus terms)
# of Documents from info sample	160771	191413	161045	62724	36016	30971	16853
# of Documents from bio sample	271907	292254	248424	86333	108940	59563	35223
# of Documents from water sample	122848	121648	127517	30264	34677	22073	17318
	555526	605315	536986	179321	179633	112607	69394
Sample terms: 840							
Average	661	721	639	213	214	134	83
sum				1168	1146	1167	1939
Overlap of documents with thesaurus terms from several EVMs				**18.27%**	**18.66%**	**11.49%**	**4.26%**

**Terms left after discarding "not found" cases:

- Subdomain EVMs: Info sample=233, Bio sample=383, Water sample=224

Experiment III Rigid Query Mode Both sample sets Sample terms:1644	Query 1 (info thesaurus Term)	Query 2 (bio thesaurus term)	Query 3 (water thesaurus term)	Query 4 (intersection info AND bio)	Query 5 (intersection bio AND water)	Query 6 (intersection info AND water)	Query 7 (intersection all thesaurus terms)
Average	635	765	693	254	232	128	84
sum				1146	1226	1200	2009
Overlap of documents with thesaurus terms from several EVMs				**22.17%**	**18.90%**	**10.63%**	**4.16%**

Experiment III Loose Query Mode M sample set*	Query 1 (info thesaurus Term)	Query 2 (bio thesaurus term)	Query 3 (water thesaurus term)	Query 4 (intersection info AND bio)	Query 5 (intersection bio AND water)	Query 6 (intersection info AND water)	Query 7 (intersection all thesaurus terms)
# of Documents from info sample	907139	953283	1571743	124487	78104	52577	33976
# of Documents from bio sample	1634836	2117609	3097587	290781	246415	148179	87877
# of Documents from water sample	1113905	1216556	1576305	107026	122378	86269	25041
	3655880	4287448	6245635	522294	446897	287025	146894
Sample terms: 804							
Average	4547	5333	7768	650	556	357	183
sum				9230	12545	11958	17465
Overlap of documents with thesaurus terms from several EVMs				**7.04%**	**4.43%**	**2.99%**	**1.05%**

*Terms left after discarding "not found" cases:

- Subdomain EVMs: Info sample=191, Bio sample=377, Water sample=236

Experiment III Loose Query Mode V sample set**	Query 1 (info thesaurus Term)	Query 2 (bio thesaurus term)	Query 3 (water thesaurus term)	Query 4 (intersection info AND bio)	Query 5 (intersection bio AND water)	Query 6 (intersection info AND water)	Query 7 (intersection all thesaurus terms)
# of Documents from info sample	1161785	1287181	1816693	122675	148572	109879	41832
# of Documents from bio sample	1900183	2261074	2749217	193290	232348	170112	76034
# of Documents from water sample	1102609	1122279	1720907	70795	103690	79625	45890
	4164577	4670534	6286817	386760	484610	359616	163756
Sample terms: 840							
Average	4958	5560	7484	460	577	428	195
sum				10058	12468	12014	17807
Overlap of documents with thesaurus terms from several EVMs				**4.58%**	**4.63%**	**3.56%**	**1.09%**

**Terms left after discarding "not found" cases:

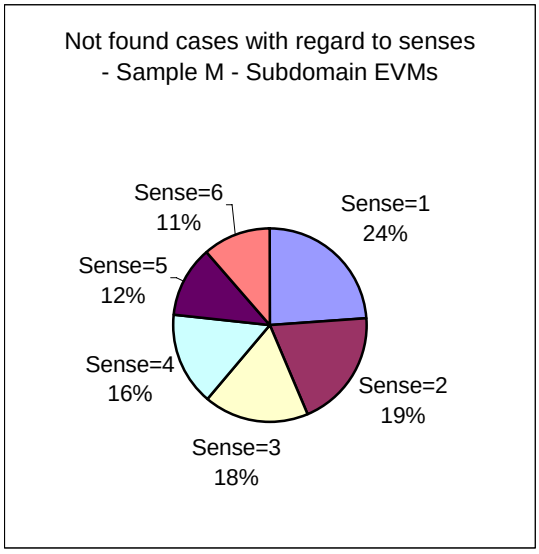
- Subdomain EVMs: Info sample=233, Bio sample=383, Water sample=224

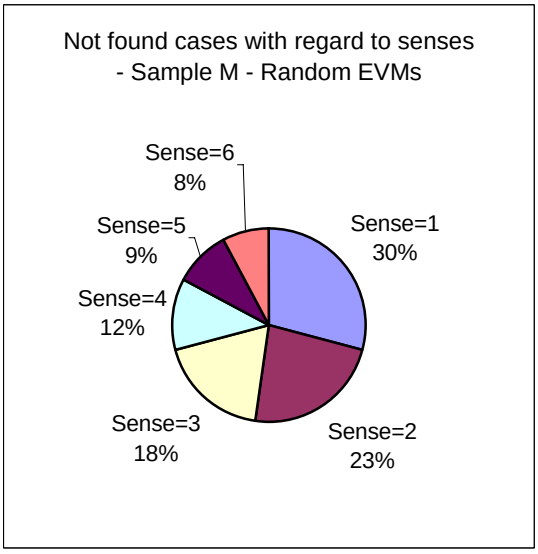
Experiment III Loose Query Mode Both sample sets Sample terms:1644	Query 1 (info thesaurus Term)	Query 2 (bio thesaurus term)	Query 3 (water thesaurus term)	Query 4 (intersection info AND bio)	Query 5 (intersection bio AND water)	Query 6 (intersection info AND water)	Query 7 (intersection all thesaurus terms)
Average	4757	5449	7623	553	567	393	189
sum				9653	12505	11987	17640
Overlap of documents with thesaurus terms from several EVMs				**5.73%**	**4.53%**	**3.28%**	**1.07%**

5. Other experiments: "Not found" cases with regard to senses

We took a detailed look at the sample terms where one of the EVMs wouldn't find an equivalent thesaurus term. We analyzed whether there is a relation between how many meanings (senses) a sample has and how likely it is that one of the EVMs doesn't find a thesaurus term. Our hypothesis stated that sample terms with a lower number of senses are more likely not to be found by one of the EVMs (reversal of experiment II).

We analyzed Sample M (600 words from the info index, bio index, and water index) both with the subdomain EVMs as with the random EVMs[9]. Our hypothesis was confirmed by both subdomain and random EVMs, although the random EVMs seem to have an even stronger tendency to miss thesaurus terms for sample terms with a lower number of senses.





[9] The random EVMs have a size of about 20,000 records.

III. Evaluation of the prediction power of Subdomain Indexes[10]

Subdomain EVMs have the task to provide a mapping from the user's natural search terms to the metadata terms (e.g. subject headings, thesaurus terms, classification codes) of a given knowledge system and help the user finding the appropriate query terms.

In the first series of experiments we analyzed how subdomain EVMs vary from a general index. In this new series of experiments we took a more scientific approach to evaluate the quality of subdomain EVMs. We measured the EVMs' prediction power in suggesting the correct and relevant metadata terms.

We tested the prediction power of EVMs by comparing the metadata terms that were originally assigned to a document and the metadata terms (in ranked order) the EVMs would predict. Although the primary function of EVMs is to predict new metadata terms for new documents (or natural search terms), we could test the quality of prediction by testing with already existing metadata terms for given documents.

In order to measure the prediction power, we defined two variables: precision and recall. Recall counts the number of retrieved relevant terms by the EVM among the number of assigned[11] terms. Precision is defined as the portion of the retrieved metadata terms (by the EVM) that is relevant. For our evaluation, we presented the precision and recall measures at different cutoff levels (cutoff levels in this case are the number of retrieved metadata terms).

> Example (by Y. Kim): At the cutoff level of one, which means taking only the top ranked terms from the suggested list of terms by EVM, if this term is one of five human indexed metadata terms, the Precision is 1.00 and the Recall is 0.20.

1. Defining the subdomain EVMs

As described in the introduction, we defined subdomain EVMs by using the INSPEC classification hierarchy going from broad categories to more specific sub categories. All sub categories are direct partitions of the broader categories.
We created the following EVMs[12]:

- A Physics consisting of 219,463 records from the INSPEC database that would have a classification code assigned starting with the letter A
- A2 Nuclear Physics consisting of 18,400 records from the INSPEC database that would have a classification code assigned starting with the letter A2
- A21 Nuclear Structure consisting of 3,133 records from the INSPEC database that would have a classification code assigned starting with the letter A21

---

[10] This work continues the efforts of Youngin Kim: Evaluation of the performance of the EVM dictionaries. June 2000. http://www.sims.berkeley.edu/research/metadata/papers/eval_desc.html
[11] We assume the assigned terms for a document are relevant.
[12] The choice of classification category was arbitrary.

- B Electrical and Electronic Engineering consisting of 145,450 records from the INSPEC database that would have a classification code assigned starting with the letter B
- B2 Components, Electron Devices and Materials consisting of 40,409 records from the INSPEC database that would have a classification code assigned starting with the letter B2
- B21 Passive circuit components consisting of 2,288 records from the INSPEC database that would have a classification code assigned starting with the letter B21

- C Computers and Control with 119,985 records from the INSPEC database that would have a classification code assigned starting with the letter C
- C5 Computer Hardware with 38,823 records from the INSPEC database that would have a classification code assigned starting with the letter C5
- C51 Circuits and Devices with 4,284 records from the INSPEC database that would have a classification code assigned starting with the letter C51

- D Information Technology with 3,896 records from the INSPEC database that would have a classification code assigned starting with the letter D

The EVMs were built by using both title and abstract of the records for indexing.
In this series of experiment we didn't experiment with varying the indexing strategy to get better results (e.g. taking only title or title and abstract for building the index; building word-based dictionaries or phrase-based dictionaries; choosing different NLP techniques for extracting noun phrases). However, all these variables could have a great impact on the prediction quality of the EVMs.

2. Training and testing the EVMs

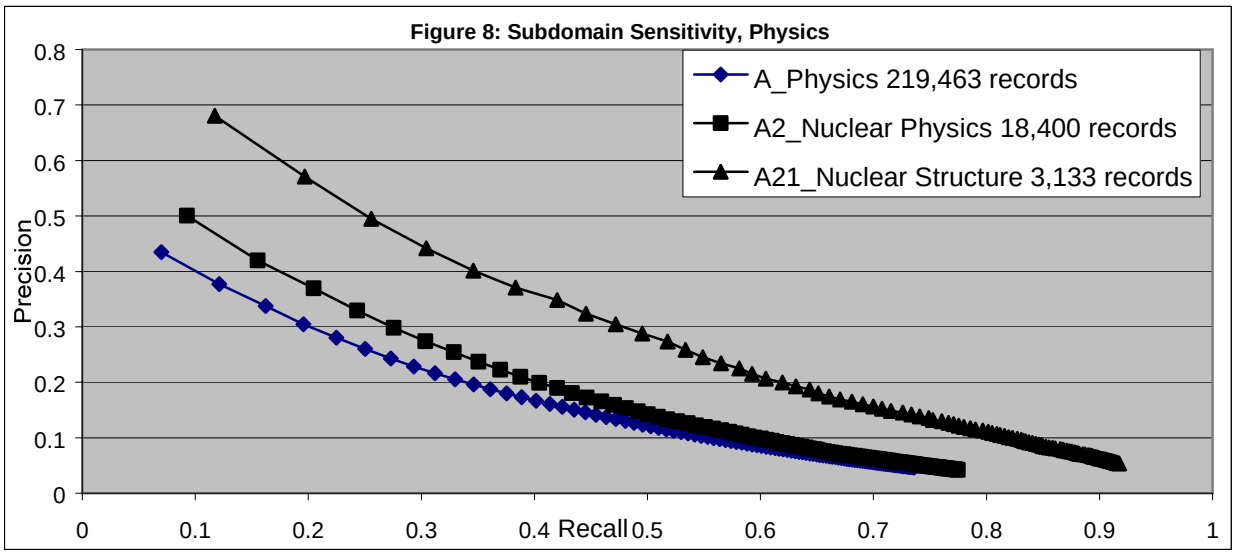
For building the EVMs, we downloaded records from the INSPEC database with the appropriate classification number (duplicates have been removed). We then divided this record pool into a training and a test set. The training set, with which the actual EVM was built, consists of 80% of the data, whereas the test set consists of 20% of the data.

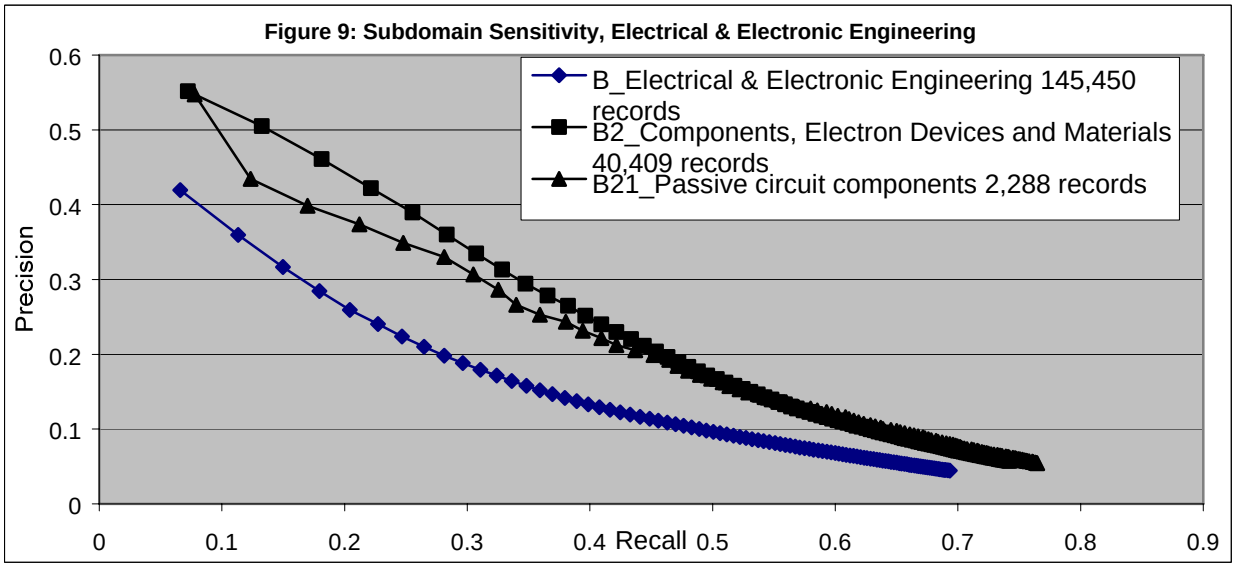
One should pay attention to the fact that the testing data (records) was always as specific as the subdomain EVM. That is, we tested very specific terms (using the title and abstract from the test record) against very specific subdomain EVMs (with only a number subject headings) and compared this to much broader defined EVMs and data sets. This method could lead to problems in later applications because we only evaluated that the EVMs are as good and precise as the test terms submitted to them. Specific EVMs and specific test data (search terms submitted to the EVMs) could still occur in some imaginable situations, e.g. if an EVM covers the content of one special academic journal and the EVM is used to predict metadata terms for new articles.

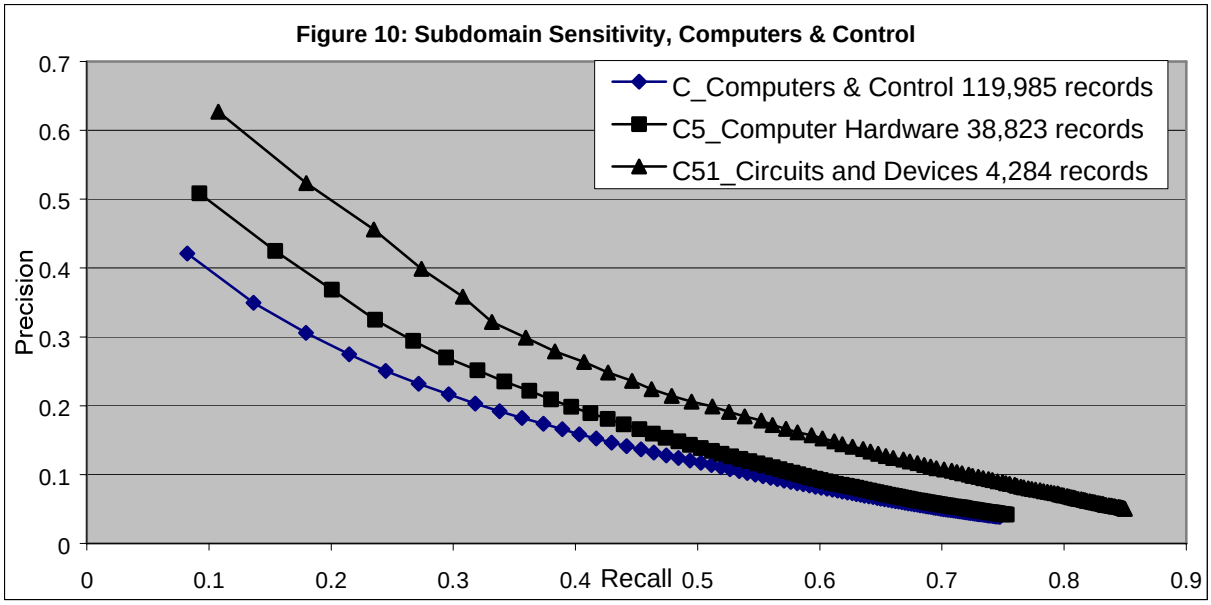
Later experiments should test the quality of EVMs with test data that varies in the specifics of search terms.

3. Evaluation results

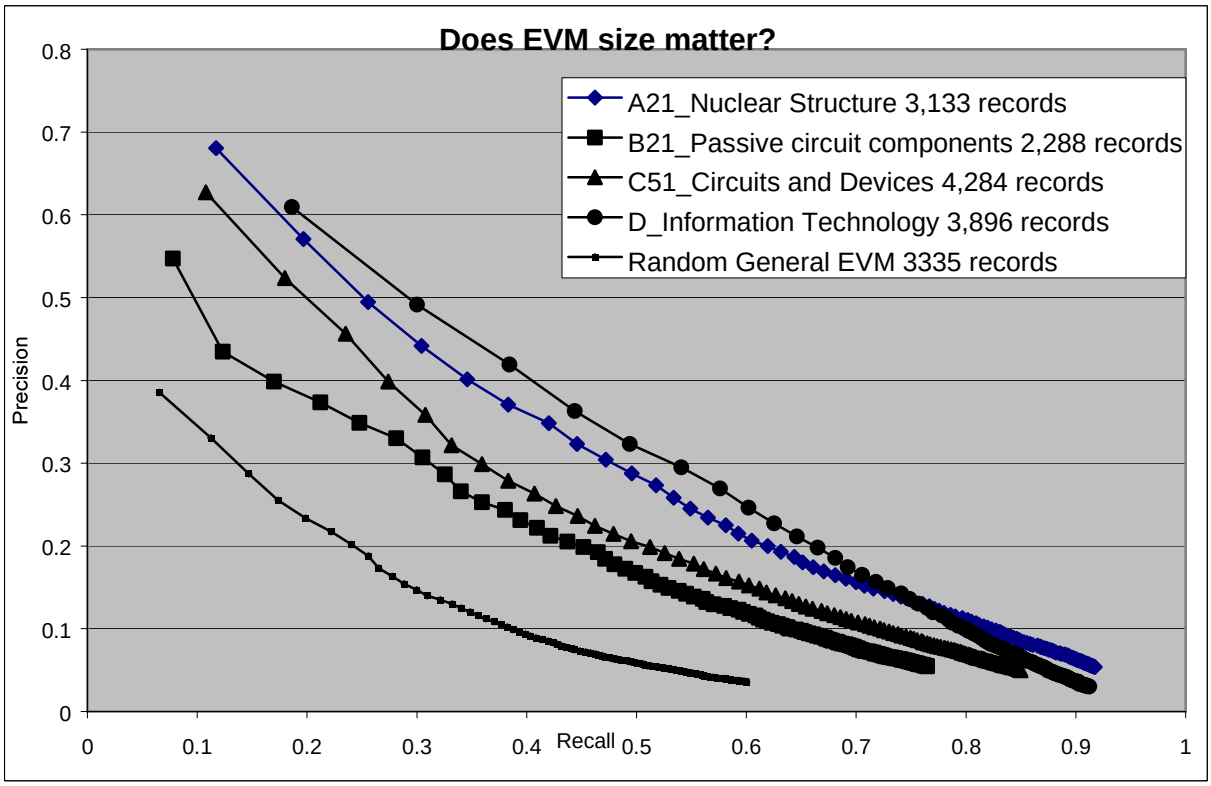
As predicted, we found that more specific subdomain EVMs (from smaller sub categories) have both better precision and recall measures than the broader defined EVMs.
However, one should consider several impact factors for this result: the specific subdomain EVMs are much smaller and have fewer unique subject headings than the broader EVMs. The vocabulary of these specific areas is probably more concise and therefore easier for an EVM to reflect in its predictions. Figures 8-10 show the results for three subject areas and its sub areas.







As a comparison, we compared the smallest EVMs with a randomly created EVM. In all cases, the subdomain EVMs performed better than the random EVM.



4. Suggestions for further experiments

To overcome the obstacle that EVMs are best in predicting subject headings for very precise search terms (or records for that matter) but the usual search terms are very broad in nature, we can use a two-stage-strategy to use EVMs for predicting correct metadata terms.
We can use a general EVM to predict whether a search term or record falls into one of the four classification categories of INSPEC (Physics, Electrical and Electronic Engineering, Computers & Control, Information Technology). Once, we associate the search terms with a more precise EVM, we can use this EVM to predict a metadata term or even can one step further to predict a more precise EVM (more specific sub category of INSPEC).

It is also important to further analyze the role of unique subject headings per EVM and their overlap. The following table gives an overview of unique subject headings for each subdomain EVM.

A	7501	B	7437	C	6240
A2	4161	B2	5513	C5	4727
A21	898	B21	2148	C51	2113
D	744				