The Role of Different Thesauri Terms and Captions in Automated Subject Classification

Koraljka Golub
KnowLib Research Group, Department of Information Technology, Lund University
Koraljka.Golub@it.lth.se

Abstract

The paper aims to explore to what degree different types of terms in Engineering Information (Ei) thesaurus and classification scheme influence automated subject classification performance. Preferred terms, their synonyms, broader, narrower, related terms, and captions are examined in combination with a stemmer and a stop-word list. The algorithm comprises string-to-string matching between words in the documents to be classified and words in term lists derived from the Ei thesaurus and classification scheme. The data collection for evaluation consists of some 35000 scientific paper abstracts from the Compendex database. A subset of the Ei thesaurus and classification scheme is used, comprising 92 classes at up to five hierarchical levels from General Engineering. The results show that preferred terms perform best, whereas captions perform worst. Stemming in most cases shows to improve performance, whereas the stop-word list does not have a significant impact.

1. Introduction

Automated subject classification (in further text: automated classification) denotes machine-based organization of information objects into topically related groups. Automated classification has been a challenging research issue for several decades now. The importance of controlled vocabularies such as thesauri in automated classification has been recognized in recent research [1, 2, 3, 4].

Vocabulary control in thesauri is achieved in several ways, out of which the following are beneficial for automated classification:

- the terms are usually noun phrases, which are content words;
- the meaning of the term is restricted to that most effective for the purposes of a particular thesaurus, which is indicated by the addition of scope notes and definitions, providing additional context for automated classification;
- three main types of relationships are displayed in a thesaurus: 1) equivalence (synonyms, lexical variants, terms treated as synonyms for general purposes); 2) hierarchical (generic, whole-part or instance relationships); 3) associative (terms that are closely related conceptually but not hierarchically and are not members of an equivalence set). In automated classification, equivalence terms allow for discovering the concepts and not just words expressing them. Hierarchies provide additional context for determining the correct sense of a term, and so do associative relationships.

The purpose of the paper is to explore to what degree different types of terms in Ei (Engineering Information) thesaurus and classification scheme [5] influence classification performance. Preferred terms, their synonyms, related, broader, narrower terms and captions are examined in combination with a stemmer and a stop-word list. The study would imply which terms with which weights to use in classification.

2. Methodology

2.1. String-matching algorithm

The algorithm searches for terms from the Ei thesaurus and classification scheme in documents to be classified. In order to do this, a term list is created, containing class captions, different thesauri terms and classes which the terms and captions denote. The list consists of triplets: term (single word, Boolean term or phrase), class which the term designates or maps to, and weight. Boolean terms consist of words that must all be present but in any order or in any distance from each other. The Boolean terms are not explicitly part of the Ei thesaurus, so they had to be created in a pre-processing step. They are considered to be those terms which contain the following strings: ‘and’ (word “and”), ‘vs.’ (short for “versus”), ‘,’ (comma), ‘;’ (semi-colon, separating different concepts in class names), ‘(‘ (parenthesis, indicating the context of a homonym), ‘:’ (colon, indicating a more specific description of the previous term in a class name), and ‘--’ (double dash, indicating heading--subheading relationship). Upper-case words from the Ei thesaurus and classification scheme are left in upper case in the term list, assuming that they are acronyms. All other words containing at least one lower-case letter are converted into lower case. Geographical names are excluded on the grounds that they are not being engineering-specific in any sense.

The following is an excerpt from the Ei thesaurus and classification scheme, based on which the excerpt from the term list (further below) was created:

From the classification scheme (captions):

931.2 Physical Properties of Gases, Liquids and Solids
…
942.1 Electric and Electronic Instruments
…
943.2 Mechanical Variables Measurements

From the thesaurus:

TM Amperometric sensors
UF Sensors--Amperometric measurements
MC 942.1
…
TM Angle measurement
UF Angular measurement
UF Mechanical variables measurement--Angles
BT Spatial variables measurement
RT Micrometers
MC 943.2
…
TM Anisotropy
NT Magnetic anisotropy
MC 931.2

TM stands for the preferred term, UF for synonym, BT for broader term, RT for related term, NT for narrower term; MC represents the main class. Below is an excerpt from the All term list (see 2.3.), as based on the above examples:

1: electric @and electronic instruments =942.1,
1: mechanical variables measurements =943.2,
1: physical properties of gases @and liquids @and solids =931.2,
1: amperometric sensors =942.1,
1: sensors @and amperometric measurements =942.1,
1: angle measurement =943.2,
1: angular measurement =943.2,
1: mechanical variables measurement @and angles =943.2,
1: spatial variables measurement =943.2,
1: micrometers =943.2,
1: anisotropy =931.2,
1: magnetic anisotropy =913.2

The algorithm looks for strings from a given term list in the document to be classified and if the string (e.g. “magnetic anisotropy” from the above list) is found, the class(es) assigned to that string in the term list (“913.2” in our example) are assigned to the document. One class can be designated by many terms, and each time the class is found, the corresponding weight (“1” in our example) is assigned to the class. The scores for each class are summed up and classes with scores above a certain cut-off (heuristically defined) can be selected as the final ones for that document. In this particular study the weight is always “1” and all the classes are assigned as the final ones. Both weights and cut-offs will be dealt with in further research, based on the results of this study.

2.2. Data collection

Data collection consists of a subset of 35166 paper titles and abstracts from the Compendex database [6], classified in the 92 selected Ei classes (cf. 2.3. Term lists). On average, 2.2 classes per document have been intellectually assigned (by humans who are experts in the subject and in indexing).

Compendex is a commercial database so the subset cannot be made available to others. However, the authors can provide records’ identification numbers on request.

2.3. Term lists

Ei classification scheme is organized into six categories which are divided into 38 subjects, which are further subdivided into 182 specific subject areas. These are further subdivided, resulting in some 800 individual classes in a five-level hierarchy.

For this study one of the six main classes was selected, together with all its subclasses: class 900 – Engineering, General. The reason for choosing this class is that it contains both natural sciences such as engineering physics, and social sciences such as engineering profession and engineering management. The latter tend to use more polysemic words than the former, and as such present a more complex challenge for automated classification.

Within the 900 main class, there are 99 subclasses, but for seven of them the number of documents in Compendex was few (less than 100), so it was decided to exclude those seven classes from the study altogether.

The table below (Table 1) shows how many different types of terms there are in the 92 classes (Total), and the average number of terms per class (Avg./class).

Table 1. The number of different types of terms

	All	BT	Ca	NT	PT	RT	ST
Total	8099	932	92	1423	1691	4378	1739
Avg./class	88	10	1	15	18	48	19

For the study, seven different term lists were created:

1. **All** – the term list containing captions, preferred terms, synonyms, related, narrower and broader terms. It consists of 8099 entries.
2. **Broader (BT)** – the term list containing only broader terms from the thesaurus. It consists of 932 entries.
3. **Captions (Ca)** – the term list containing only captions from the list of classification codes. “Caption” stands for the term translating the class number (e.g. class number “900” has caption “Engineering, General”). This list consists of 92 entries.
4. **Narrower (NT)** – the term list containing only narrower terms from the thesaurus. It consists of 1423 entries.
5. **Preferred (PT)** – the term list containing only preferred terms from the thesaurus. It consists of 1691 entries.
6. **Related (RT)** – the term list containing only related terms from the thesaurus. It consists of 4378 entries.
7. **Synonyms (ST)** – the term list containing only non-preferred terms from the thesaurus. It consists of 1739 entries.

2.4. Stop-word list and stemming

Terms and captions in the Ei thesaurus and classification scheme can also contain words which are frequently used in many contexts and as such are not very indicative of any document’s topicality (e.g. the word “general” in the Ei class caption “Engineering, General”). The stop-word list used contained 429 words, and was taken from [7]. For stemming, the Porter Stemming Algorithm was used [8]. The stop-word list was applied to the term lists, and stemming to the term lists as well as documents.

2.5. Evaluation methodology

Assuming that intellectually assigned classes in the data collection are correct, evaluation in this study is based on comparison of automatically derived classes against the intellectually assigned ones. The subset of the Ei thesaurus and classification scheme used in the experiment comprises 92 classes at five hierarchical levels. These 92 classes are all related to each other – often there is only a small topical difference between them. The topical relatedness is expressed in numbers representing the classes – the more initial digits any two classes have in common, the more related they are. Thus, comparing the classes at only the first few digits instead of all the five (each representing one hierarchical level), would also make sense. Still, the evaluation in this study is conducted based on all the five different levels, i.e., an automatically assigned class is considered correct only when it is exactly the same as an intellectually assigned class for the same document.

Apart from the standard micro-averaged and macro-averaged precision, recall and F1 measures ([9], p.33), the results are compared based on the number of documents that got assigned at least one class (Clas. doc. in Tables 2, 3 and 4), and the average number of classes assigned to each document (Avg. nbr. clas. in Tables 2, 3 and 4). There are about 2.2 classes intellectually assigned per document, and the aim of automated classification is to achieve similar.

3. Experimental results

3.1. Averaged results for all the classes

Table 2. No stop-word list, no stemming

	All	BT	Ca	NT	PT	RT	ST
Clas. doc. %	**96.8**	85.9	16.6	56.3	74.2	94.6	39.4
Avg. clas. nbr.	16.1	6.8	0.2	1.0	**1.7**	11.2	0.7
Macroa. P	0.11	0.11	0.43	0.29	**0.48**	0.12	0.37
Macroa. R	**0.54**	0.24	0.05	0.07	0.22	0.37	0.11
Microa. P	0.07	0.08	**0.43**	0.21	0.30	0.08	0.33
Microa. R	**0.54**	0.26	0.04	0.10	0.23	0.41	0.10
Macroa. F1	0.15	0.10	0.06	0.08	**0.22**	0.13	0.12
Microa. F1	0.13	0.12	0.07	0.13	**0.26**	0.13	0.15
Avg. F1s	0.14	0.11	0.07	0.10	**0.24**	0.13	0.14

As seen from Tables 2, 3 and 4, the best performance measured as mean F1s (Avg. F1s) has the Preferred term list, and the worst one the Captions list. As seen from Table 3, and by comparing the mean F1s of Tables 2 and 3, stemming showed to be beneficial in four out of the seven different term lists: Captions, Narrower, Preferred, and Synonyms. In All and Related lists the F1 performance got worse due to too much lowered precision. Table 4 shows impact of the stop-word list, and in comparison to Table 2, the mean of F1s improved for Narrower and Preferred terms. For other terms the stop-word list didn’t do much of a difference since the thesaurus contains only content words, as seen from the last two rows in Table 4. The last row (Stop-w. %) shows the percentage of stop-words in all the terms on a list. In all lists apart from the shortest list (Captions), less than 10% are stop-words. Captions list has only 92 terms (Table 1) but the terms in this list are mostly longer than terms in other lists. This is due to the fact that captions' original function in a classification scheme is to describe well what the corresponding class number stands for, and not to be a distinct term.

Table 3. No stop-word list, stemming

	All	BT	Ca	NT	PT	RT	ST
Clas. doc. %	**99.4**	97.2	28.6	87.3	95.6	99.1	71.3
Avg. nbr. clas	28.3	12.8	0.4	**2.6**	4.2	19.9	1.6
Macroa. P	0.09	0.09	**0.42**	0.27	0.40	0.10	0.33
Macroa. R	**0.72**	0.38	0.07	0.14	0.36	0.54	0.16
Microa. P	0.06	0.06	**0.36**	0.15	0.20	0.07	0.22
Microa. R	**0.73**	0.38	0.06	0.19	0.38	0.59	0.16
Macroa. F1	0.13	0.10	0.08	0.11	**0.27**	0.13	0.15
Microa. F1	0.10	0.11	0.10	0.17	**0.26**	0.12	0.18
Avg. F1s	0.11	0.11	0.09	0.14	**0.26**	0.12	0.17

Table 4. Stop-word list, no stemming

	All	BT	Ca	NT	PT	RT	ST
Clas. doc. %	**97.8**	86.5	16.5	70.0	81.1	95.6	44.6
Avg. nbr. clas	17.5	7.1	0.2	1.4	**2.1**	12.2	0.9
Macroa. P	0.11	0.11	0.42	0.30	**0.47**	0.12	0.36
Macroa. R	**0.56**	0.24	0.05	0.08	0.23	0.38	0.12
Microa. P	0.07	0.08	**0.42**	0.22	0.29	0.08	0.27
Microa. R	**0.59**	0.26	0.04	0.14	0.28	0.42	0.11
Macroa. F1	0.15	0.10	0.06	0.09	**0.22**	0.13	0.13
Microa. F1	0.13	0.12	0.07	0.17	**0.28**	0.13	0.16
Avg. F1s	0.14	0.11	0.07	0.13	**0.25**	0.13	0.14
Stop-w. nbr.	473	39	13	131	156	259	101
Stop-w. %	5.8	4.2	14.1	9.2	9.2	5.9	1.1

Concerning the number of classes per document that get automatically assigned, when using Captions less than one class is assigned on average even when stemming is applied; Narrower and Synonyms improve with stemming, close to our aim of 2,2 classes that have been intellectually assigned. The most appropriate number of classes get assigned when Preferred terms are used with stop-words. Based on All, Broader and Related lists, too many classes get assigned, but that could be dealt with in the future by introducing cut-offs (cf. last paragraph of 2.1.).

The results are similar for the number of documents that get classified: in all the three tables the lowest number of documents gets classified using Captions (less than 30% even when stemming is applied), then using Synonyms and Narrower terms in their best case (71% and 87% respectively). When looking at Table 1, such results could be expected for Captions since only one term designates a class. On the other hand, Preferred and Synonyms lists have similar number of terms, but Preferred performs almost twice as good when stemming is not applied. This reflects the fact that preferred terms in the Ei thesaurus occur more frequently in the documents than their synonyms.

By comparing results in Table 5 with numbers of terms in each term list (Table 1), we can see that only a certain number of terms is found in the documents being classified, ranging from 29% for All (when no stop-words and no stemming are used), to 74% for Captions (when stemming is used).

Table 5. Number of found terms from term lists

	All	BT	Ca	NT	PT	RT	ST
No stop-words, no stemming	2348	387	62	651	823	1453	553
No stop-words, stemming	2730	413	68	789	960	1632	701
Stop-words, no stemming	2359	385	62	657	823	1455	558

3.2. Individual classes

As seen earlier from Table 1, each class is on average designated by 88 terms, ranging from 1 to 756 terms per class. The majority of terms are related terms, followed by synonyms and preferred terms.

Table 6 lists top performing classes using the All term list, no stemming and no stop-word list (their F1, either micro-averaged or macro-averaged, is above 0.30). The table also shows the number of each type of terms per class. We can see that the sole number of terms designating a class does not seem to be proportional to the performance. Moreover, these best-performing classes do not have a similar distribution of types of terms designating them, i.e. the percentage of certain term types does not seem to be directly related to performance.

Table 7 lists worst-performing classes using the All term list, no stemming and no stop-word list (their F1, either micro-averaged or macro-averaged, is 0.03 or less). As it is the case with the best-performing classes, the worst-performing classes do not have a similar number of classes designating them, neither do they have a similar distribution of types of terms designating them.

Table 8 compares performance of the same worst-performing classes as Table 7, in regards to involving or not the stop-word list and stemming. The differences are very small, but stemming has a negative effect in 8 out of 10 cases, whereas stop-word list improves in two cases and worsens in two others.

Table 6. Top performing classes and number of terms

Class	All	BT	Ca	NT	PT	RT	ST	F1
941.1	24	4	1	1	3	12	3	0.32
933.1.1	135	13	1	7	21	54	39	0.32
931.3	510	42	1	52	93	234	88	0.33
921.5	58	6	1	8	9	25	9	0.33
932.2.1	50	3	1	4	9	16	17	0.35
944.4	11	1	1	0	1	8	0	0.38
903	28	1	1	3	5	17	1	0.38
933.3	5	0	1	0	1	2	1	0.39
903.3	54	7	1	4	9	17	16	0.42
913.4.3	1	0	1	0	0	0	0	0.54
Total	876	77	10	79	151	385	174	n/a
Avg./class	88	8	1	8	15	39	17	n/a

Table 7. Worst-performing classes and number of terms

Class	All	BT	Ca	NT	PT	RT	ST	F1
911.5	13	2	1	0	2	6	2	0.02
941.4	88	10	1	9	18	38	12	0.02
913	31	5	1	1	6	18	0	0.02
913.3.1	25	1	1	0	2	19	2	0.02
941	126	12	1	18	18	55	22	0.03
922	34	4	1	3	6	14	6	0.03
912.1	185	9	1	42	19	96	18	0.03
933	185	19	1	38	24	86	17	0.03
943.3	360	43	1	41	73	141	61	0.03
911.3	17	2	1	0	2	8	4	0.03
Total	1064	107	10	152	170	481	144	n/a
Avg./class	106	11	1	15	17	48	14	n/a

Table 8. F1 of worst-performing classes in relation to stemming and stop-words

	F1		
Class	No stemming, no stop-word list	Stemming, no stop-word list	No stemming, stop-word list
911.5	0.017	0.011	0.017
941.4	0.022	0.022	0.022
913	0.022	0.021	0.023
913.3.1	0.023	0.016	0.030
941	0.025	0.021	0.020
922	0.027	0.019	0.027
912.1	0.027	0.020	0.027
933	0.028	0.021	0.025
943.3	0.029	0.029	0.029
911.3	0.030	0.028	0.030

4. Concluding remarks

The majority of classes is found when using the All term list and stemming: micro-averaged recall is 73% (Table 3). The remaining 27% of classes were not found because the words in the term list designating the classes did not exist in the text of the documents to be classified.

In the study, no weighting or cut-offs were applied, but will be experimented with in the future. This study implies that all types of terms should be used for a term list (All) in order to achieve best recall, but that higher weights could be given to preferred terms, captions and synonyms, as the latter yield highest precision. Stemming seems useful for achieving higher recall, and could be balanced by introducing weights for stemmed terms. Stop-word list could be applied to captions, narrower and preferred terms.

Acknowledgment

This work was supported by the IST Programme of the European Community under ALVIS (IST-1-002068-STP). The author thanks Anders Ardö whose detailed comments helped improve the paper.

References

[1] T. Koch, A. Ardö, “Automatic classification”, DESIRE II D3.6a, Overview of Results, 2000. Available: http://www.lub.lu.se/desire/DESIRE36a-overview.html.

[2] S. L. Bang, J. D. Yang, and H. J. Yang, ”Hierarchical document categorization with k-NN and concept-based thesauri”, Information Processing and Management, 2006, 42, pp. 387–406.

[3] O. Medelyan, and I. Witten, “Thesaurus based automatic keyphrase indexing”, in: Proc. of the JCDL 2006, pp. 296–297.

[4] P. J. Garcés, J. A. Olivas, and F. P. Romero, “Concept-matching IR systems versus word-matching information retrieval systems: Considering fuzzy interrelations for indexing Web pages”, JASIS&T 2006, 57(4), pp. 564–576.

[5] Ei thesaurus, J. Milstead, Ed. 2nd ed. Castle Point on the Hudson Hoboken: Engineering Information, 1995.

[6] Compendex database. Available: http://www.engineeringvillage2.org/.

[7] Onix text retrieval toolkit: Stop word list 1. Available: http://www.lextek.com/manuals/onix/stopwords1.html

[8] M. Porter. Porter stemming algorithm. Available: http://www.tartarus.org/martin/PorterStemmer/.

[9] F. Sebastiani, “Machine learning in automated text categorization”, ACM Computing Surveys, 2002, 34(1), pp. 1–47.