PRINTER-FRIENDLY FORMAT Return to Article

D-Lib Magazine

September/October 2013
Volume 19, Number 9/10

Bringing Order to Digital Libraries: From Keyphrase Extraction to Index Term Assignment

Nicolai Erbs
Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt

Iryna Gurevych
Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt and German Institute for Educational Research and Educational Information

Marc Rittberger
German Institute for Educational Research and Educational Information

Point of contact for this article: Nicolai Erbs, erbs@ukp.informatik.tu-darmstadt.de

doi:10.1045/september2013-erbs

Abstract

Collections of topically related documents held by digital libraries are valuable resources for users; however, as collections grow, it becomes more difficult to search them for specific information. Structure needs to be introduced to facilitate searching. Assigning index terms is helpful, but it is a tedious task even for professional indexers, requiring knowledge about the collection in general, and the document in particular. Automatic index term assignment (ITA) is considered to be a great improvement. In this paper we present a hybrid approach to index term assignment, using a combination of keyphrase extraction and multi-label classification. Keyphrase extraction efficiently assigns infrequently used index terms, while multi-label classification assigns frequently used index terms. We compare results to other state-of-the-art approaches for related tasks. The assigned index terms allow for a clustering of the document collection. Using hybrid and individual approaches, we evaluate a dataset consisting of German educational documents that was created by professional indexers, and is the first one with German data that allows estimating performance of ITA on languages other than English.

Keywords: Keyphrase, Keyphrase Extraction, Automatic Index Term Assignment, Multi-label Classification

1. Introduction

Topically related documents are collected in digital libraries to provide a rich resource for users who are interested in certain topics. A topically coherent collection is a good starting point for a focused search. However, as a collection grows it gets more complicated to find information. Some kind of structure needs to be introduced to facilitate searching. If manually created overview pages that partition the collection into separate topic areas are not frequently updated, they become rather misleading, because they disregard new documents. An alternative to facilitate searching is index terms.

Index terms are helpful as they (i) provide a summary of the document [34], (ii) improve searching [29], and (iii) allow a dynamic partitioning of topics [7]. Assigning index terms is a tedious task because it requires knowledge about the collection in general, and the document in particular. Collection knowledge is important because a good index term highlights a specific subtopic of a coarse collection-wide topic. Document knowledge is important because a good index term is a summary of the document's text. This problem is not limited to English document collections. In other languages, e.g. German, other problems arise. Thesauri which are available for English are not available in every language and less training data may be available.

In digital libraries, the task of assigning index terms can be performed by professional indexers. But even professional indexers do not always follow a strict scheme. Some index terms are assigned to a large part of the whole collection, while others are assigned to a single document only. Over the course of time, a list of frequently used coordinate index terms is composed by the indexers. Occasionally, non-listed index terms are assigned to a document. As a result, some index terms (infrequent) are best suitable for searching and some (frequent) are best suitable for clustering documents.

Automatic index term assignment is a great improvement for digital libraries by making indexing documents easier and hence making finding relevant documents faster. However, the variation of index terms makes it hard to represent different types of index terms in a single model. Therefore, we analyze the characteristics of index terms and present approaches for assigning each type. We propose a hybrid approach to better model the manually assigned index terms which allows both automatic assignment of index terms for searching and for providing suggestions to professional indexers. It is a novel approach, since we combine keyphrase extraction and multi-label classification directly. The direct combination allows selecting the best performing algorithms for each of the approaches individually.

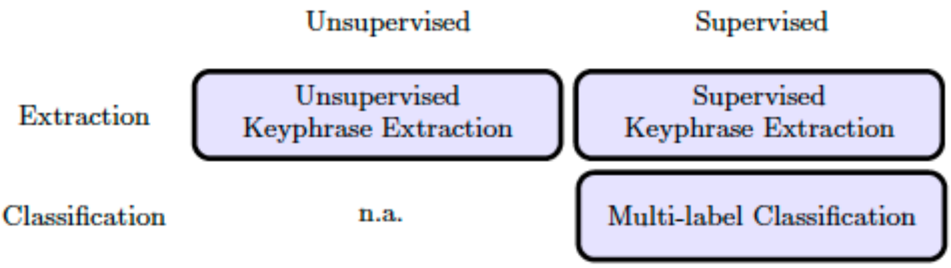
We also present a classification of approaches that can be used for index term assignment, introduce a dataset consisting of German documents, and evaluate the presented approaches on this dataset. We evaluate the approaches individually and as a hybrid approach. We conclude with a summary of our findings and a discussion of future work.

2. Classification of Approaches

We define index term assignment as a generic term for approaches assigning phrases to a document for indexing and searching. Newman *et al.* [21] defines index term assignment as "the task of automatically determining terms to include in a literary index for a document collection". This literary index may contain up to 100 index terms and should give a reference to its occurrence. Newman states that assigning index terms requires knowledge of the whole document collection, while keyphrases are assigned based on a single document. Csomai *et al.* [5] apply approaches from keyphrase extraction to the task of back-of-the-book indexing, which is highly related to index term assignment.[1] They incorporate knowledge resources such as Wikipedia to compute the "keyphraseness" of index terms. Hulth [9] trains a supervised system for keyphrase extraction using linguistic features such as part-of-speech patterns.

The closest work to ours is from Medelyan *et al.* [18]. They apply keyphrase extraction to the task of automatic tagging of documents [18], which is equivalent to index term assignment. We extend their work by incorporating approaches from multi-label classification.

Figure 1 provides a coarse classification of approaches that can be applied to index term assignment. Keyphrase extraction and multi-label classification approaches are distinguished in the vertical axes. Additionally, Figure 1 separates unsupervised and supervised approaches. Unsupervised approaches do not require any training data, while supervised approaches consist of a training phase to create a model and a testing or evaluation phase where the trained model is applied to new documents.



	Unsupervised	Supervised
Extraction	Unsupervised Keyphrase Extraction	Supervised Keyphrase Extraction
Classification	n.a.	Multi-label Classification

Figure 1: Overview of approaches for index term assignment

2. 1 Keyphrase Extraction

Keyphrase extraction approaches rank terms from the document according to a metric. Hence, they are based on terms that appear in the document text. Keyphrases are a good means for searching a huge document collection. Typically, only a small fraction of documents share the same keyphrases which allow for fast searching of the whole collection [29].

Salton *et al*. [25] lay the foundation of keyphrase extraction with the introduction of the **tf—idf** metric. It relates the term frequency inside a document with the number of documents in the collection containing the term. Equation 1 shows the basic formularization of tf—idf. In this formula $f(t, d)$ is the frequency of term t in document d, $|D|$ is the number of documents and $|d \in D : t \in d|$ is the number of documents mentioning term t.

$$(1) \quad \text{tf–idf}(t,d) = f(t,d) \cdot \log \frac{|D|}{|d \in D : t \in d|}$$

The first term measures the importance of the term inside the current document; the second term measures the *distinctiveness* of the term, i.e. terms that are infrequently used in other documents are very distinctive. Terms that appear frequently inside a document but infrequently in other documents of the collection have a high tf—idf value. Common terms like stopwords (*and, but*) have a high frequency in all documents and thus have a lower tf—idf value. Terms with a medium to high tf—idf value are taken as keyphrases.

Tomokiyo *et al*. [31] use language models to compute *informativeness*, i.e. terms with high information content and phraseness, i.e. terms that appear often as a multiword. Csomai *et al*. [4] adopt this idea and use tf—idf values as a measure for *informativeness*. The phrase **Soccer World Cup** has high tf—idf values (*informativeness*) and the words are often used in this combination (*phraseness*), hence making it a good keyphrase.

Different configurations of the tf—idf metric have shown to perform better in some scenarios [17]. As a modification to tf—idf, the inverse document frequency can be replaced with the term frequency in a background corpus *D'* [24], shown in Equation 2.

$$\text{tf–idf}_{\text{background}}(t,d) = f(t,d) \cdot \log \sum_{d' \in D'} f(t,d') \quad (2)$$

The modified inverse document frequency is more like a background term frequency. Instead of counting the frequency in the corpus itself, the frequency in a — usually larger — background corpus[2] is computed. Equation 3 applies the inverse document frequency without taking the logarithm of the inverse document frequency and Equation 4 sets the inverse document frequency to a constant value of 1. The latter configuration corresponds to text frequency.

$$\text{tf–idf}_{\text{normal}}(t,d) = f(t,d) \cdot \frac{|D|}{|d \in D : t \in d|} \quad (3)$$

$$\text{tf–idf}_{\text{constant}}(t,d) = f(t,d) \cdot 1 \quad (4)$$

Mihalcea *et al*. [20] introduce the unsupervised graph-based approach **TextRank** to extract keyphrases: In the graph, candidates for keyphrases are used as the nodes, and an edge is added if two keyphrase candidates co-occur in a certain context window (e.g. 3 words left or right of the anchor candidate, or in the same sentence as the anchor candidate) in the document. The weight of the edge is defined as the number of co-occurrences. A graph centrality measure, e.g. PageRank [2], is then used to rank the nodes in the graph. The highest ranked nodes are then taken as keyphrases. This approach is corpus-independent; no information from external resources is taken into account.

In comparison to unsupervised approaches, supervised approaches are able to combine the previously introduced metrics. In a training phase a model specific to the dataset is learned and in the second phase applied to new documents. Supervised approaches apply machine learning algorithms, e.g. decision trees [35] or Naïve Bayes [37], to solve this problem.[3]

As features for supervised approaches, tf—idf values and cooccurrence (TextRank) information can be taken. The position of a candidate is also a good feature as keyphrase might appear earlier in the document. Machine learning allows for using many features for which their importance is learned in a training phase. Using more features may improve results further. Part-of-speech information is valuable, as nouns are more likely keyphrases than prepositions [9]. Additionally, acronym identification techniques can be applied as they may also be good keyphrases [22]. Supervised approaches often outperform unsupervised systems [12] for keyphrase extraction, as they are better able to incorporate many aspects of the corresponding document and external resources.

2.1.1 Controlled vocabulary

An additional source of information is a thesaurus. It is a list of previously collected terms which indexers often use for assigning index terms to documents. Medelyan *et al*. [19] states that the usage of a controlled vocabulary "eliminates the occurrence of meaningless or obviously incorrect phrases". They specify keyphrase extraction when using such a controlled vocabulary as index term extraction. Lopez *et al*.[15] use the existence of a term in domain-specific vocabulary as a feature.

To the best of our knowledge, the distinction between keyphrases and index terms is not clear-cut and the presented approaches can be applied to both index term assignment and keyphrase extraction. As both provide summaries for a document and improve searching, we apply keyphrase extraction approaches to the task of index term assignment. However, two similar documents might not share any keyphrases due the great variety of possible keyphrases. This leads to sparse document clusters which do not help users in finding topically coherent documents. In the following section, we present an approach to assign frequently appearing index terms shared by many documents.

2.2 Multi-label classification

Multi-label classification assigns labels from a predefined label set to a document. Documents can be clustered by their labels which allows creating overview pages and browsing through the collection [11]. These labels are similar to tags or categories [27] and the set of all possible labels is also referred to as a tag set. Instead of extracting keyphrases from the document only, any label from the label set can be assigned [14].

Multi-label classification [16] first learns a classification model for each of the labels based on features, e.g. words. If many training documents labeled with `university education` contain the word `professor`, a classifier will label a new document containing the same word accordingly in the second step. This is the case, even if the document does not explicitly contain the phrase `university education`.

Figure 1 above lists no approaches for unsupervised classification. However, there exist unsupervised document classification approaches which we do not present in this work, e.g. [28]. They merely perform document clustering, while we focus on assigning index terms for documents.

3. Dataset

Our data consists of peer-reviewed articles, dissertations, and books from the educational domain published by researchers. We extract all documents from the database dump of peDOCS and select all German documents (91% of all documents). Documents span all topics related to education, e.g. historical and general education, pedagogy of media, and environment. Hence, it is heterogeneous in terms of style, length, and level of detail.

Professional indexers assign all index terms in the peDOCS documents as postcoordinate index terms. This is different from index terms in other datasets, where gold standard index terms are usually author-assigned [10, 22]. In our case, the indexers follow certain guidelines and apply them to every document in the collection consistently.

Table 1 lists characteristics of index term assignment datasets including peDOCS. Previously presented datasets [10, 36, 22] are collected using English documents, our dataset contains German documents. INSPEC contains abstracts of scientific papers, DUC documents are newspaper articles, and SP (Scientific Publications) are publications in the domain of computer sciences. Thus, INSPEC contains on average the shortest and SP the longest documents. Documents in peDOCS are also scientific publications and on average longer than documents in the SP dataset. The average number of index terms is comparable on all three English datasets (ranging from 8.08 to 9.64). Documents in peDOCS have on average slightly more index terms (11.57 index terms).

Dataset	peDOCS	INSPEC [10]	DUC [36]	SP [22]
# of documents	3,424	2,000	301	134
Language	German	English	English	English
Ø length (# of tokens)	14,531	139	903	8,492
Ø index terms per document	11.57	9.64	8.08	8.31

Table 1: Corpus statistics of index term assignment datasets

In addition to the documents of the peDOCS dataset, professional indexers have two thesauri from which they can select index terms. Indexers are not restricted to index terms from these thesauri; however, they were constructed by constantly adding frequently assigned index terms. This leads to the construction of a "core" thesaurus with 974 terms and an "extended" thesaurus of 8,487 terms. The core thesaurus captures 41.8% and the extended thesaurus captures 83% of all assigned index terms.

We further compute the ratio of index terms that appear in the peDOCS dataset. Our analysis on peDOCS shows that 67.3% of the index terms appear in the document text they are assigned to. Keyphrase extraction approaches are limited to assigning index terms that appear in the text, hence making 67.3% the upper bound.

Those index terms that do not appear in the document may be assigned by applying multi-label classification approaches. Thus, we analyze the frequency of index terms in peDOCS. Figure 2 shows the frequency distribution of index terms.

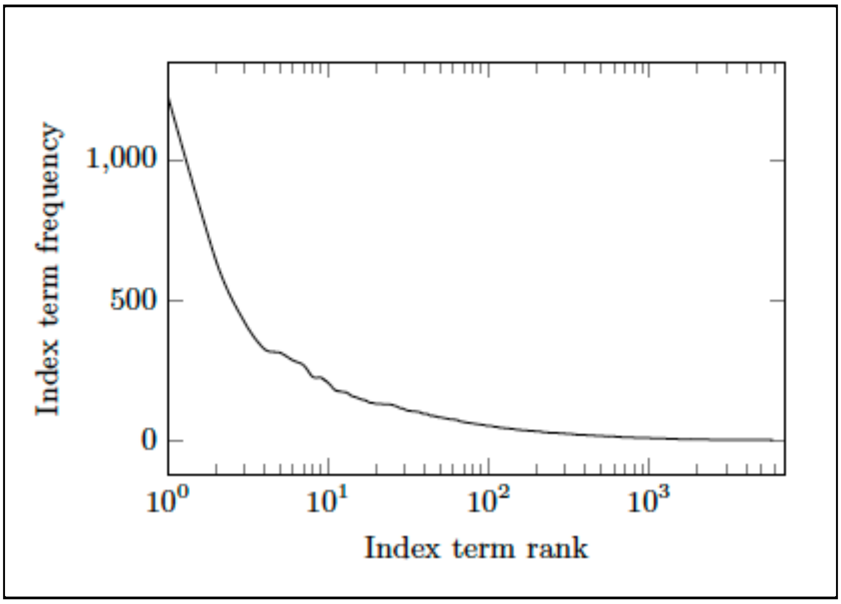


Figure 2: Frequency distribution of index terms following a power-law distribution

We first rank index terms according to their frequency and then plot the resulting distribution. It follows a power-law distribution which proves that few index terms are used for many documents and many index terms are used only once. Examples of the most frequently used index terms are *Deutschland* (Eng: *Germany*), *Schule* (Eng: *school*), and *Schüler* (Eng: *pupil*). These rather general terms do not capture the specific topic of the document. As the collection contains many example documents labeled with theses index terms, we apply multi-label classification to assign these index terms in new documents.

In this section we presented the peDOCS dataset and compared it to other datasets. peDOCS is the largest dataset for index term assignment in terms of number of documents, average document length, and average number of index terms per document. It is also the first one with German data and indexed by professional indexers.

4. Experimental Results

We apply keyphrase extraction and multi-label classification approaches on the peDOCS dataset. Our hybrid approach is comparable to the behavior of indexers: we assign frequently used index terms and then assign additional index terms extracted from the document. We model our hybrid approach by combining the highest ranked keyphrases and labels. Therefore, we first analyze the results of the individual approaches in the following sections. We use a randomly sampled subset of 2,400 documents as our test data.

4.1 Keyphrase Extraction Experiments

We evaluate unsupervised and supervised keyphrase extraction approaches using precision, recall, and R-precision (all micro-averaged) as evaluation metrics.[4] Precision is defined as the ratio of correctly extracted keyphrases in the identified keyphrase list $\mathrm{P} = \frac{|K \cap G|}{|K|}$ where *K* is the set of automatically extracted keyphrases and *G* the set of manually assigned keyphrases (gold keyphrases). Recall is defined as the ratio of correctly extracted keyphrases that are assigned to the document $\mathrm{R} = \frac{|K \cap G|}{|G|}$. R-precision is defined as precision when the number of extracted keyphrases is limited by the number of manually assigned keyphrases [38]. We lemmatize[5] extracted and manually assigned keyphrases to map different forms of the same word.

Table 2 displays the results for keyphrase extraction. We provide information for the upper bound for all three evaluation metrics. It is possible to reach 100% in terms of precision as all extracted keyphrases can be correct. However, recall is limited to 67.3% as only keyphrases that appear in the text can be extracted. The same applies to R-precision. The results are grouped into unsupervised and supervised approaches. We present results for a simple position metric (the earlier the keyphrase appears in the text, the higher it is ranked), different configurations for the tf—idf metric, and TextRank as another state-of-the-art approach (cf. Section 2.1). Supervised approaches use Naïve Bayes to learn a model using features which are themselves unsupervised approaches. We apply 10-fold cross-validation to evaluate supervised approaches.

Type	Approach	Precision	Recall	R-precision
—	*Upper bound*	*100%*	*67.3%*	*67.3%*
Unsupervised	Position	8.3%	11.2%	5.6%
	tf—idf	7.5%	10.1%	6.9%
	tf—idf$_{\text{background}}$	9.2%	12.4%	11.7%
	tf—idf$_{\text{normal}}$	0.4%	0.5%	0.3%
	tf—idf$_{\text{constant}}$	11.6%	15.5%	10.2%
	TextRank	9.7%	13.1%	8.7%
Supervised	tf—idf + Position	10.6%	14.2%	**11.9%**
(Naïve Bayes)	tf—idf$_{\text{background}}$ + Position	**11.9%**	**16.0%**	11.7%

Table 2: Results for keyphrase extraction approaches on peDOCS

Using the position as a weight for index terms yields good results as scientific papers usually start with an abstract mentioning the most relevant terms. Our evaluation of different modifications of tf—idf shows that using the tf—idf$_{\text{constant}}$ as defined in Equation 4 performs best in terms of precision and recall. tf—idf$_{\text{background}}$ (cf. Equation 2) yields best unsupervised results in terms of R-precision. The weaker performance of classic tf—idf might be due to many keyphrases that are used throughout the dataset which leads to a high document frequency and thus to a low tf—idf value. TextRank yields results comparable to tf—idf$_{\text{background}}$ in terms of precision and recall, but lower results in terms of R-precision.

Supervised approaches using a variation of tf—idf and the position of the candidate further improve results. Using background frequency (tf—idf$_{\text{background}}$) instead of document frequency (tf—idf) yields better results. As expected, supervised approaches perform better than the corresponding unsupervised approaches. The model is able to learn characteristics of the peDOCS dataset and we expect that increasing the size of the dataset yields even better results.

We observe a different ranking of approaches depending on the evaluation metric. Precision and recall are based on the best 10 extracted keyphrases. R-precision, on the contrary, considers as many extracted keyphrases as there are manually assigned. Hence, the number of manually assigned keyphrases and the order of extracted keyphrase influence results: R-precision achieves better results if the ranking within the top-10 extracted keyphrases is better.

Keyphrase extraction approaches allow for extracting many keyphrases with a good recall and a reasonable precision. The extracted keyphrases can directly be used as index terms. In addition to keyphrase extraction we evaluate with multilabel classification approaches in the following section to enhance precision for index term assignment.

4.1.1 Comparison to results on English data

Table 3 compares results on the German dataset peDOCS with the three English datasets INSPEC, DUC, and SP. It shows results obtained with TextRank [20] and the tf—idf approach. The highest precision is obtained using tf—idf on the INSPEC dataset which is also the one with the on average shortest documents. The highest recall is obtained using TextRank on the SP dataset, and tf—idf on the DUC dataset performs best in terms of R-precision. A direct comparison of the results is hard because preprocessing (e.g. POS tagging) for different languages does not perform equally well. Overall, results are very low which might be due to skipping a very common filtering of manually assigned keyphrases not appearing in the document. The presented results show that keyphrase extraction on German data is — similar to keyphrases extraction on English data — a hard task.

Dataset	Approach	Precision	Recall	R-precision
peDOCS	TextRank	9.7%	13.1%	8.7%
	tf-idf	7.5%	10.1%	6.9%
INSPEC	TextRank	11.7%	10.5%	6.4%
	tf-idf	**12.5%**	11.3%	6.2%
DUC	TextRank	5.5%	12.2%	6.9%
	tf-idf	8.9%	11.1%	**8.9%**
SP	TextRank	8.3%	**13.8%**	8.6%
	tf-idf	4.3%	5.1%	4.5%

Table 3: Results of keyphrase extraction across all datasets

4.1.2 Controlled vocabulary

Previous work [19, 15] states that domain-specific or controlled vocabularies further improve performance of keyphrase extraction. Thus, we use the described thesauri for peDOCS (cf. Section 3) as a filter for our extracted keyphrases.

Table 4 shows results of the best performing keyphrase extraction system (tf—idf$_{background}$) on the peDOCS dataset. Not using any filter leads to an R-precision of 11.7%. Using the extended list with 8,487 terms reduces the potential coverage to 83% but R-precision rises to 26.5%. A further reduction to the core thesaurus does not increase results. Instead, filtering with the core thesaurus returns an R-precision of 22.2%.

Vocabulary	# Terms	Coverage	R-precision
None	39,616	100%	11.7%
Extended list	8,487	83.0%	**26.5%**
Core list	974	41.8%	22.2%

Table 4: Results of keyphrase extraction using controlled vocabulary with tf—idf$_{background}$

Using controlled vocabularies improves keyphrase extraction approaches. However, as a thesaurus is not always available, we limit our approaches to the more likely case of not using any thesaurus.

4.2 Multi-label Classification Experiments

As previously shown in Table 2, index terms are not equally distributed. Some index terms are used only once, while a few are used very frequently. These frequently used index terms can be assigned with multi-label classification. Instead of using a thesaurus, we use the most frequent index terms as labels for classification. We introduce the parameter *n* as the size of our label set. More index terms can be covered if *n* is set to a higher value but on average fewer examples will be available for each label. Examples are documents for which an indexer has assigned labels. Classification algorithms require positive (documents with a specific label) and negative (documents without this label) examples to learn a model; for less frequent index terms there are not enough positive examples.

We evaluate results for the multi-label classification approach under identical conditions as done for keyphrase extraction. We compare classified labels to manually assigned index terms and measure results in terms of precision, recall, and R-precision.[6] We use the open source software tool Mulan [32] based on WEKA [7] and apply cross-validation to avoid leaking information from the learning to the evaluation phase. We use the top-500 n-grams[7] from the dataset as features. We use two frequently used classification approaches: support vector machines (SVM)[8] [3] and decision trees (J48)[9] [23].

The label set size *n* determines the upper bound for assigning index terms (cf. Table 5). Extending the label set allows a higher recall (increase from 11.3% for 10 labels to 31.6% for 200 labels). However, precision decreases for a larger label set size, especially, in case of SVM for which precision drops from 50.1% (10 labels) to 32.2% (200 labels). A label set size of 200 is a good trade-off between precision and recall. Although larger label sets are possible, we limit the label set size to 200 as computation time increases with size.[10]

Both classification algorithms perform almost on par. J48 provides better results in terms of recall and R-precision and SVM in terms of precision. For label set size of 10, SVM reaches a high precision of 50.1%.

Overall, results in Table 5 show that multi-label classification assigns index terms (labels) with higher precision but lower recall compared to keyphrase extraction. The low results in terms of recall and R-precision are due to a lower number of classified labels: Recall is limited if less than ten labels are classified.

Label set *n*	Algorithm	Precision	Recall	R-precision
10	*Upper bound*	*100%*	*11.3%*	*11.3%*
	J48	35.8%	4.3%	5.8%
	SVM	**50.1%**	3.2%	5.1%
20	*Upper bound*	*100%*	*15.4%*	*15.4%*
	J48	30.9%	5.5%	**6.6%**
	SVM	40.6%	3.9%	5.5%
50	*Upper bound*	*100%*	*23.4%*	*23.4%*
	J48	30.5%	**6.3%**	6.5%
	SVM	33.7%	5.0%	5.8%
200	*Upper bound*	*100%*	*31.6%*	*31.6%*
	J48	33.0%	6.1%	**6.6%**
	SVM	32.2%	6.0%	6.3%

Table 5: Results for multi-label classification approaches

4.3 A Combination of Approaches

Previously, we presented keyphrase extraction approaches that assign index terms appearing in the text and multi-label classification approaches that assign frequently used index terms. As the purpose of our system is to assign both types of index terms, we present a hybrid system, combining the strengths of both approaches. Like a professional indexer, our system assigns index terms appearing in the document itself and index terms appearing frequently in the whole collection. We combine the highest ranked index terms from supervised keyphrase extraction and multi-label classification. Terms from both approaches are added to a list with at most ten index terms. In case one approach returns less than five index terms (which is often the case for multi-label classification), more index terms from the other approach are added.

Table 6 displays the best performing approaches from Table 2 and Table 5. Additionally, it displays a combination of the supervised keyphrase extraction and multi-label classification approach. Supervised keyphrase extraction performs better than unsupervised keyphrase extraction and multilabel classification obtains the best precision, but lowest recall and R-precision. A combination of supervised keyphrase extraction and multi-label classification retains a high precision, while keeping a fair recall. The hybrid system achieves the highest values for recall and R-precision.

#	Type	Approach	Precision	Recall	R-precision
1	Unsupervised KE	tf-idf$_{constant}$	11.6%	15.5%	10.2%
2	Supervised KE	tf-idf$_{background}$ + Position	11.9%	16.0%	11.7%
3	Multi-label Cl.	J48 (200 labels)	**33.0%**	6.1%	6.6%
	Hybrid system	2nd + 3rd approach	20.0%	**17.9%**	**14.4%**

Table 6: Results for combinations of approaches

Table 7 provides manually assigned index terms for an example document[11] and index terms assigned with supervised keyphrase extraction, multi-label classification and our hybrid system (combination of both). In total 21 index terms are manually assigned to this documents (we listed the first eleven). Some index terms are very similar, e.g. *Statistik* (Eng: *statistics*) is the general term for *Bildungsstatistik* (Eng: *educational statistics*). Supervised keyphrase extraction returns a weighted list which is cut-off after ten index terms. Multi-label classification assigns four index terms for this document.

Manually Assigned Index Terms	2: Supervised Keyphrase Extraction	3: Multi-label Classification	Hybrid System
Bildungsforschung	**Hochschule** (0.96)	**Statistik** (1.0)	**Hochschule**
Wissenschaft	Personal (0.94)	Bildungspolitik (0.75)	**Statistik**
Hochschullehrerin	**Wissenschaft** (0.94)	Schule (0.40)	Personal
Berufung	Bildungsplanung (0.93)	Berufsbildung (0.33)	Bildungspolitik
Professur	Fortschreibung (0.93)		**Wissenschaft**
Hochschule	**Frau** (0.92)		Schule
Forschung	Kulturwissenschaft (0.90)		Bildungsplanung
Wissenschaftler	Frauenanteil (0.89)		Berufsbildung
Frau	Datenmaterial (0.87)		Fortschreibung
Bildungsstatistik	Hausberufung (0.87)		**Frau**
Statistik	...		
...			

Table 7: Manually and automatically assigned index terms for an example document. Correctly assigned index terms are marked bold and scores for each assigned index term are in parentheses if available.

Supervised keyphrase extraction successfully assigns three index terms within the top-10 list and multi-label classification correctly assigns the highest ranked index term. The combination of both approaches yields best results (correctly assigning four index terms). We observe several near misses of index terms, e.g. *Hausberufung* (Eng: *internal appointment*) is a near miss for *Berufung* (Eng: *appointment*). Additionally, we observe that most of the assigned index terms not appearing in the assigned list, are still good index terms for the document, e.g. *Frauenanteil* (Eng: *percentage of women*) and *Bildungspolitik* (Eng: *education policy*). Rather than counting matches of index terms, an extrinsic evaluation of their usefulness could provide better insights into a system's quality.

5. Summary and Future Work

In this paper, we analyzed the peDOCS dataset for index term assignment in detail and compared it to other datasets. peDOCS is the first dataset consisting of German documents and with index terms that are manually assigned by professional indexers. Its document count and average number of index terms per document is larger than for other datasets.

We presented approaches for index term assignment and analyzed their strengths and shortcomings. Keyphrase extraction assigns many potential index terms but is restricted to index terms that appear in the document. Multi-label classification assigns index terms with high precision but is limited to a predefined set of labels. Our experimental results on peDOCS showed that keyphrase extraction assigns index terms with low precision and higher recall, while multi-label classification obtains higher precision and low recall. A hybrid system combining both approaches yields better results in terms of recall and R-precision. Our error analysis shows that a hybrid system successfully assigns a large proportion of manually assigned index terms and provides further index terms not reflected by manually assigned index terms.

For future work, we plan to verify our observations in a user study. Our hypothesis is that index terms assigned by automatic approaches are comparable to manually assigned index terms and useful for indexing and clustering in digital libraries. Additionally, incorporating semantic resources may improve automatic evaluation for index term assignment by bridging the gap between related index terms. A semantic evaluation may use synonymy relations in WordNet [6] to match different index terms. Additionally, matches of index terms can be weighted using similarity measures [1].

Acknowledgements

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, by the Klaus Tschira Foundation under project No. 00.133.2008, and by the German Federal Ministry of Education and Research (BMBF) within the context of the Software Campus project open window under grant No. 01IS12054. The authors assume responsibility for the content. We thank Yevgen Chebotar, Shuo Yang, and Artem Vovk for their valuable contributions and the we thank the anonymous reviewers for their helpful comments.

Notes

[1] The order of words in back-of-the-book indexes are often altered to be an inverted index [4].

[2] We use the German Reference Corpus *DeReKo* from 2012 [13].

[3] For an overview of machine learning algorithms, see [13].

[4] We measure precision and recall at 10 extracted keyphrases.

[5] We use the TreeTagger trained on German data [26].

[6] We do not report the accuracy for each of the labels as we are interested in the overall performance for index term assignment.

[7] We use unigrams, bigrams and trigrams.

[8] Using RakEL [33] as meta algorithm.

[9] Using BRkNN [30] as meta algorithm.

[10] 10-fold cross-validation takes about 18 hours on a workstation (quad-core processor).

[11] Title: *Chancengleichheit in Wissenschaft und Forschung* (Eng: *Equal Opportunities in Science and Research*)

References

[1] D. Bär, T. Zesch, and I. Gurevych. A Reflective View on Text Similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 515—520, Hissar, Bulgaria, 2011.

[2] S. Brin and L. Page. The Anatomy of a Large-scale Hypertextual Web Search Engine. In *Proceedings of the Seventh International Conference on World Wide Web*, pages 107—117, 1998.

[3] C. Cortes and V. Vapnik. Support-vector Networks. *Machine Learning*, 297:273—297, 1995.

[4] A. Csomai and R. Mihalcea. Investigations in Unsupervised Back-of-the-book Indexing. *Proceedings of the Florida Artificial Intelligence Research Society*, (Hulth):211—216, 2007.

[5] A. Csomai and R. Mihalcea. Linguistically motivated features for enhanced back-of-the-book indexing, pages 932—940, 2008.

[6] C. Fellbaum. WordNet. *Theory and Applications of Ontology: Computer Applications*, pages 231—243, 2010.

[7] C. Gutwin, G. Paynter, I. Witten, C. Nevill-Manning, and E. Frank. Improving Browsing in Digital Libraries with Keyphrase Indexes. *Decision Support Systems*, 27(1-2):81—104, Nov. 1999.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: an Update. *SIGKDD Explorations*, 11(1):10—18, 2009.

[9] A. Hulth. Improved Automatic Keyword Extraction given more Linguistic Knowledge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 216—223, 2003.

[10] A. Hulth. Enhancing Linguistically Oriented Automatic Keyword Extraction. In *Proceedings of HLT-NAACL: Short Papers*, pages 17—20, 2004.

[11] R. Jäschke and L. Marinho. Tag Recommendations in Folksonomies. In *Knowledge Discovery in Databases: PKDD*, pages 506—514, 2007. http://doi.org/10.1007/978-3-540-74976-9_52

[12] S. Kim, O. Medelyan, M. Kan, and T. Baldwin. Semeval-2010 Task 5: Automatic Keyphrase

Extraction from Scientific Articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21—26, 2010.

[13] M. Kupietz and C. Belica. The German Reference Corpus DeReKo: a Primordial Sample for Linguistic Research. In *Proceedings of the 7th Conference on International Language Resources and Evaluation*, pages 1848—1854, 2010.

[14] M. Lipczak. Tag Recommendation for Folksonomies Oriented towards Individual Users. *ECML PKDD Discovery Challenge*, 84, 2008.

[15] P. Lopez and L. Romary. HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 248—251, 2010.

[16] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski. An Extensive Experimental Comparison of Methods for Multi-label Learning. *Pattern Recognition*, 45:3084—3104, 2012.

[17] C. D. Manning, P. Raghavan, and H. Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 2008.

[18] O. Medelyan, E. Frank, and I. H. Witten. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1318—1327, 2009.

[19] O. Medelyan and I. H. Witten. Thesaurus based Automatic Keyphrase Indexing. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 296—297, 2006.

[20] R. Mihalcea and P. Tarau. TextRank: Bringing Order into Texts. In *Proceedings of Empirical Methods for Natural Language Processing*, pages 404—411, 2004.

[21] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100—108. Association for Computational Linguistics, 2010.

[22] T. Nguyen and M.-Y. Kan. Keyphrase Extraction in Scientific Publications. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, pages 317—326. Springer, 2007.

[23] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1992. http://doi.org/10.1007/BF00993309

[24] P. Rayson and R. Garside. Comparing Corpora Using Frequency Profiling. In *Proceedings of the Workshop on Comparing Corpora*, pages 1—6, 2000.

[25] G. Salton and C. Buckley. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513—523, 1988.

[26] H. Schmid. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, volume 21, pages 1—9, 1995. http://doi.org/10.1007/978-94-017-2390-9_2

[27] F. Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computing Surveys (CSUR)*, 34(1):1—47, 2002.

[28] N. Slonim, N. Friedman, and N. Tishby. Unsupervised Document Classification Using Sequential Information Maximization. In *Proceedings of the 25th International Conference on Research and Development in Information Retrieval*, pages 129—136, 2002.

[29] M. Song, I. Y. Song, R. B. Allen, and Z. Obradovic. Keyphrase Extraction-based Query Expansion in Digital Libraries. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 202—209, 2006.

[30] E. Spyromitros, G. Tsoumakas, and I. Vlahavas. An Empirical Study of Lazy Multilabel Classification Algorithms. In *Artificial Intelligence: Theories, Models and Applications*, pages 401—406. Springer, 2008. http://doi.org/10.1007/978-3-540-87881-0_40

[31] T. Tomokiyo and M. Hurst. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment*. Volume 18, pages 33—40. Association for Computational Linguistics, 2003.

[32] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining Multi-label Data. In *Data Mining and Knowledge Discovery Handbook*, 2nd edition, pages 667—685. Springer, 2010.

[33] G. Tsoumakas, I. Katakis, and I. Vlahavas. Random k-labelsets: An Ensemble Method for Multilabel Classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1079—1089, 2011. http://doi.org/10.1007978-3-540-74958-5_38

[34] S. Tucker and S. Whittaker. Have A Say Over What You See: Evaluating Interactive Compression Techniques. In *Proceedings of the 2009 International Conference on Intelligent User Interfaces*, pages 37—46, 2009.

[35] P. D. Turney. Learning Algorithms for Keyphrase Extraction. *Information Retrieval*, 2(4):303—336, 2000. http://doi.org/10.1023/A:1009976227802

[36] X. Wan and J. Xiao. Single Document Keyphrase Extraction using Neighborhood Knowledge. *Proceedings of AAAI*, pages 855—860, 2008.

[37] I. Witten, G. Paynter, and E. Frank. KEA: Practical Automatic Keyphrase Extraction. In *Proceedings of the 4th ACM Conference on Digital Libraries*, pages 254—255, 1999.

[38] T. Zesch and I. Gurevych. Approximate Matching for Evaluating Keyphrase Extraction. In *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing*, pages 484—489, 2009.

About the Authors

Nicolai Erbs is a doctoral researcher in natural language processing at the UKP Lab in Darmstadt, Germany. His research focuses on keyphrase extraction, named entity disambiguation, and processing of multilingual user-generated data. Mr. Erbs received a Master of Science degree in Physics from the TU Darmstadt in 2010. In 2012 he was a visiting researcher at the IXA Group at the Basque University in Spain. He received a project grant from the German Federal Ministry of Education and Research as part of the Software Campus, for which he investigates techniques to foster self-directed learning through natural language processing.

Iryna Gurevych is Lichtenberg-Professor in Computer Science at the Technische Universität Darmstadt and Director of the Information Center for Education at the German Institute for Educational Research (DIPF) in Frankfurt, Germany. She has extensive knowledge of natural language processing techniques and their innovative applications to knowledge discovery in scientific literature. Ms. Gurevych has published over 150 papers, among them papers in ACL, NAACL, EACL, ASIS&T.

Marc Rittberger studied Physics and Information Sciences and and was a researcher in Konstanz, Düsseldorf and Genève. He is director of the Information Center for Education at the German Institute for Educational Research (DIPF) since 2005 and currently Deputy Executive Director of DIPF.

Copyright © 2013 Nicolai Erbs, Iryna Gurevych and Marc Rittberger

PRINTER-FRIENDLY FORMAT Return to Article