International Journal on Digital Libraries manuscript No.
(will be inserted by the editor)

Fusion Architectures for Automatic Subject Indexing under Concept Drift

Analysis and Empirical Results on Short Texts

Martin Toepfer · Christin Seifert

Received: date / Accepted: date

Abstract Indexing documents with controlled vocabularies enables a wealth of semantic applications for digital libraries. Due to the rapid growth of scientific publications, machine learning based methods are required that assign subject descriptors automatically. While stability of generative processes behind the underlying data is often assumed tacitly, it is being violated in practice. Addressing this problem, this article studies explicit and implicit concept drift, that is, settings with new descriptor terms and new types of documents, respectively. First, the existence of concept drift in automatic subject indexing is discussed in detail and demonstrated by example. Subsequently, architectures for automatic indexing are analysed in this regard, highlighting individual strengths and weaknesses. The results of the theoretical analysis justify research on fusion of different indexing approaches with special consideration on information sharing among descriptors. Experimental results on titles and author keywords in the domain of economics underline the relevance of the fusion methodology, especially under concept drift. Fusion approaches outperformed non-fusion strategies on the tested data sets, which comprised shifts in priors of descriptors as well as covariates. These findings can help researchers and practitioners in digital libraries to choose appropriate methods for automatic subject indexing, as is finally shown by a recent case study.

Keywords automatic subject indexing · concept drift · meta-learning · multi-label classification · short texts

M. Toepfer
ZBW – Leibniz Information Centre for Economics, Düsternbrooker Weg 120, 24105 Kiel, Germany
E-mail: m.toepfer@zbw.eu

C. Seifert*
University of Passau, Innstraße 43, 94032 Passau, Germany
University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands
E-mail: c.seifert@utwente.nl
*The article was mainly written while C. Seifert was affiliated at the University of Passau.

1 Introduction

Access to literature is best supported by subject indexes constructed using domain-specific controlled vocabularies and thesauri. Such structured representations enable semantic queries and discovery even across language barriers, and they provide features for services like literature recommendation systems. Due to the rapid growth of scientific publications [2], scalability of the indexing process has become essential, making automatic subject indexing a key technology for digital libraries.

Compared to manual indexing, automatic indexing faces several challenges: First, legal restrictions might prevent the usage of publication full-text and/or abstracts, which leads to little information available to the indexing approach and thus decreases performance [6]. Second, the distribution of concepts in the training data set can be very skewed and some concepts might not appear at all [25]. This is particularly likely for thesauri containing several thousands of concepts, as for example, the EuroVoc vocabulary[1], Medical Subject Headings (MeSH)[2], AGROVOC[3] in the agricultural domain, or the STW Thesaurus for Economics (STW)[4]. Concepts with little or no document coverage have to be

[1] `www.eurovoc.europa.eu`, accessed 28.11.2017
[2] `www.nlm.nih.gov/mesh`, accessed 28.11.2017
[3] `www.fao.org/agrovoc`, accessed 28.11.2017
[4] `www.zbw.eu/en/stw-info`, accessed 28.11.2017

either excluded [25] or require carefully designed feature spaces and concept representations for so-called zero-shot learning approaches [24]. Third, terminology in documents and controlled vocabularies might differ from each other, or they may change over time. For instance, the STW is permanently updated to reflect changes in economics literature [9]. Consider phrases like "online advertising" or "smartphone" that emerged since 1990 [22], just to give an example. Thus, indexing approaches must be capable of adapting to concept drift [8], i.e. to vanishing or emerging concepts and new types of documents containing unseen terms.

Research in the field of automatic indexing can be broadly categorized into lexical approaches and associative approaches. *Lexical approaches* like, for example, KEA++ [20] build upon knowledge provided by thesauri to find candidate concepts. Subsequently candidates are ranked and selected according to their relevance. As pointed out by Medelyan and Witten [20], this procedure requires only hundreds of training examples in total. But it comes at a cost. Lexical approaches will fail on missing candidates and incomplete vocabulary. In terms of Pouliquen et al. [25], a natural language thesaurus is required which nearly exhaustively covers the terminology of the domain. Construction and maintenance of such lexical resources is costly, thus many thesauri provide concepts but lack vocabulary entry terms, especially if multiple languages are involved. In this case, *associative approaches* may be more appropriate. They rely on associations between terms and concepts that are derived from large intellectually indexed document collections [25]. Especially, a multitude of supervised learning approaches has been proposed driven by advances in artificial intelligence and machine learning where indexing has been regarded as a multi-label learning task [10]. In essence, these approaches involve training classifiers for each concept of a thesaurus. Encouraging results have been reported in different domains, for instance, in medicine [13, 36], agriculture [17], legal texts [18], or economics [11]. Such approaches enable automatic indexing with conceptual thesauri [25] when a lot of professionally indexed examples are available, however, they do not scale well in terms of necessary training data [20]. Researchers attempted to combine elements from associative and lexical approaches aiming to alleviate their disadvantages (e.g., [13, 5, 23, 28]) with *fusion architectures*, meta-learning, or zero-shot learning techniques. Nevertheless, fusion architectures are still an exception rather than the rule, no thorough analysis of single and fusion architectures has been performed yet, and fusion can be realized in different ways. In this paper, we aim for a detailed analysis of associative, lexical and fusion architectures supported by an empirical study of a new fusion approach in the domain of economics that especially considers dynamics in terms and concepts.

Performance of automatic subject indexing systems is influenced by several factors, raising questions about generalizability. Attempts to conduct large-scale experimentation and to empirically determine successful configurations [11] provide important feedback for practitioners and researchers, but they should be supplemented by analytical justifications if possible. Recently, there have also been concerns about just concentrating on better results on standard benchmark data and how techniques like deep learning have been applied in the field of computational linguistics. For instance, Manning wanted to "encourage everyone to think about problems, architectures, cognitive science, and the details of human language, how it is learned, processed, and how it changes, rather than just chasing state-of-the-art numbers on a benchmark task" [19, p. 706]. Following this advice, we aim to gather knowledge about reasonable architectures for automatic subject indexing systems, understanding their success and pitfalls. In particular regarding zero-shot learning, humans can still outperform data-demanding applications of deep learning [16]. For further investigation of these topics, this article especially considers learning and classification performance with respect to events that are caused by differences in distributions between training and test data.

In this article we address the following research questions, regarding documents in economics:

RQ1: How can implicit and explicit concept drift be determined in a data set and how can both be visualized?

RQ2: What are advantages and disadvantages of current indexing approaches? Which combinations could potentially improve indexing performance?

RQ3: Does combination of statistical associative and lexical approaches improve indexing performance, especially for settings with concept drift?

This article is an extended version of previous work [32].[5] Among others, it adds the detailed discussion on concept drift (answering RQ1), and additionally provides results of a case study in which professional indexers rated the results of fusion approaches (answering RQ3). Although our work and the used data sets focuses on

[5]

economic literature, we provide detailed theoretical discussions, that may help researchers and practitioners in other domains.

After a recap of related work (Section 2) and the subject indexing task (Section 3), we focus on concept drift (Section 4), introduce basic terminology and demonstrate its appearance in a practical setting. Subsequently we analyse existing indexing architectures in detail in Section 5. Based on the theoretical analysis we then describe our approach to a fusion architecture that combines lexical and associative characteristics in Section 6. Results of experiments on documents from the economic science domain are presented in Section 7. Section 8 reports on recent experience with bringing a fusion system to practice, which directs to future work (Section 9). Finally, Section 10 concludes the work.

2 Related Work

We review related work in automatic subject indexing with respect to statistical associative and lexical indexing approaches and subject indexing in the economic domain. Further, we discuss different ensemble and fusion approaches as well as zero-shot learning scenarios and concept drift.

2.1 Statistical Associative Approaches

Ferber [6] developed a system with a linear *associative* model that was based on titles (*short text*) and co-occurrence data between words and descriptors. He reported encouraging results but noted that titles were sometimes insufficient and that it was unclear if the co-occurrence approach generalizes to different domains. Pouliquen et al. [25] investigated indexing with EuroVoc and found that only approximately one third of all training documents contained labels of the corresponding descriptors verbatim. For this reason, they distinguished between *conceptual thesauri* like EuroVoc and *natural language thesauri*. Because the former lack vocabulary terms for dictionary matching approaches, they proposed to determine associate terms, that is, statistically related terms, for descriptors with a statistical system similar to Ferber. Pouliquen et al. determined these associate lists by log-likelihood and then assigned descriptors by a linear combination of three similarity measures. They were able to apply the approach successfully to different languages, however, it frequently assigned descriptors that were semantically similar but wrong. Loza and Fürnkranz [18] automatically indexed legal documents of the EU using three different multi-label classification approaches based on perceptrons: binary relevance, multiclass multi-label perceptrons, and multi-label pairwise perceptrons. Pairwise classification into almost 4,000 classes of the EuroVoc vocabulary required almost 8,000,000 perceptrons. As a consequence, they had to solve severe scalability issues. Wilbur et al. [36] showed on a subset of MeSH headings that training with stochastic gradient descent (SGD) applied to support vector machines (SVM) performed well with a fixed number of iterations for ranking and prediction. SGD-SVM produced better results than several methods, including MTI, kNN-based systems, and a learning-to-rank approach. Lauser and Hotho [17] indexed full-text documents in the agricultural domain with binary SVMs. They explored different modes (*add*, *replace*, *only*) to encode background knowledge from an ontology. These modes modified the feature vectors by adding, replacing or restricting features to ontology concepts, respectively. Relations between concepts were used up to a maximum concept integration depth. Some configurations yielded slight increases in precision, however, they were not significant. The rationale behind their approach was to represent documents of the same subject areas more similarly. Section 5 includes a comparison of lexical approaches to strategies that combine term features with concept features in statistical associative systems in the aforementioned way.

2.2 Lexical Approaches

Lexical automatic indexing approaches try to recognize the terms that are stored for each concept in the controlled vocabulary. Subsequently, matches are ranked. Just to give an example, automatic subject indexing can be realized as a variant of *keyphrase extraction*, which aims to determine the most relevant phrases of full-texts to describe their contents. As shown by Medelyan et al. [20], slight modifications to a supervised keyphrase extraction system [7], can be used for subject indexing when a thesaurus with appropriate labels is available. Their system, named KEA++, filters the full-text by matching of pseudo-phrases, that is, conflated versions of the documents' terms and a controlled vocabulary's labels. Candidates are subsequently ranked and selected by a classifier. They especially pointed out that opposed to text categorization approaches, it already performs well with little training data.

2.3 Indexing in the Domain of Economics

Große-Bölting et al. [11] evaluated several configurations for semantic document annotation of documents

on three data sets. Different annotation candidate extraction and activation methods were combined with one of two kinds of selection approaches: top-k and k-nearest-neighbors (kNN). While top-k only assigns phrases that are part of the controlled vocabulary, kNN can only assign concepts for which training instances exist. Their best results on a data set in *economics* with 62,924 documents (full-text) were produced by kNN ($k = 1$; micro-averaged F_1 value of .39). By contrast to the implementations compared by Große-Bölting et al., the fusion approach investigated in this article combines lexical as well as statistical associative knowledge, while still maintaining the capability to assign precise concepts for which no training instances are available.

2.4 Ensembles and Fusion

Erbs et al. [5] pointed out differences between keyphrase extraction and multi-label classification (MLC), underlined certain advantages of MLC like detecting hidden synonyms and keyphrase extraction, and presented an approach which combines them, adding keyphrase extraction results to the list of terms returned by MLC. SVMs and decision trees were used for MLC and different configurations with TF-IDF for keyphrase extraction. They focused on full-text representations of German documents in the educational domain in their evaluation. The combined system reached 20% precision and 17.9% recall. Different from our approach, they investigated keyphrase extraction, that is, index terms were part of the documents' terms (uncontrolled vocabulary).

Nam et al. [23] aimed to predict previously unseen non-terminal concepts in concept hierarchies. They proposed a joint space of instances and concepts, using hierarchical information and concept co-occurrence patterns. Experiments were conducted on two data sets. The authors stated that the regularization approach was effective to predict previously unseen classes when the tree-structure of classes is known and not complex. A pre-training strategy was proposed that empirically improved results even on large sets of classes. Recently, Sappadla et al. [28] proposed an approach in order to exploit similarities between concept labels and document terms. To predict known concepts, they used a supervised method (binary relevance), whereas unknown concepts were predicted using label word similarity by word embeddings based on Wikipedia. They evaluated their system on three fulltext data sets. The number of classes to predict were 90 (Reuters), 45 (MEDICAL), and 201 (EURLEX). The average sizes of assigned labels were 1.23, 1.24, and 2.21, respectively. These figures are close to 1, hence, close to single-label multi-class classification. Experimentally, they were able to show advantages of their approach against a supervised baseline. When labels were removed by their frequency from the evaluation, using similarity knowledge led to higher macro-averaged metrics.

Research on automatic subject indexing has been very active in the (bio-)medical domain. Notably, the work of Jimeno-Yepes et al. [13] combined different subsystems to index MEDLINE citations with medical subject headings (MeSH). Their baseline system was the Medical Text Indexer (MTI) which was compared to several machine learning approaches (Naïve Bayes, Rocchio, AdaBoostM1, Voting) and dictionary matching on titles and titles and abstracts. They learned a mapping-table that determined which method is to be used for each MeSH heading (MH). In order to select the best method, they applied significance tests. They found that more than 23,000 MHs were best indexed by MTI, while machine learning approaches were chosen for 2,712 MHs. Combinations of machine learning methods have also been applied for categorization of genomics documents by Aronson et al. [1] who used the term *fusion* in the sense of *ensemble* or *stacking* [37, 31]. Please note that this notion differs from fusion architectures as understood in this paper (cf. Section 6). Ensemble methods like voting have often been applied only on top of several statistical associative approaches (e.g. [1]). Approaches that have applied statistical as well as lexically based methods have typically chosen one method per concept [13, 28]. In the remainder of this article, we apply fusion approaches that aim to unite individual skills. Our rationale is that if methods predict concepts differently but reliably, the union of them fully leverages their complementarity.

2.5 Zero-shot Learning

The problem of predicting previously unseen classes has been studied in other domains before, in so-called zero-shot learning settings. For instance, Palatucci et al. [24] presented an approach that uses a knowledge base to decode neural activity. As they pointed out, it is desirable to treat classes not separately from each other, but to create representations that apply to many, also unseen classes. Regarding one-shot classification and generation of visual concepts, Lake et al. [16] demonstrated improvements over deep learning approaches. How automatic subject indexing can be best realized in this regard is a current research question. Recently, some aspects have been targeted, like the aforementioned prediction of non-terminals [23] or using label embeddings for settings that are close to single-label classification [28].

2.6 Concept Drift

In general, automatic subject indexing under *concept drift* has not been been studied comprehensively, although some authors have referred to it. Tsoumakas et al. [35], for example, reported that they aimed to minimize differences between training and test data for their system when participating in an indexing challenge. For adaptation, they created focused data sets, restricting training data to the journals tested in the challenge, and the most recent documents. Different topics that are associated with concept drift in automatic subject indexing have been studied [26, 15, 12, 8, 14, 30], as explained in Section 4 in detail.

3 Subject Indexing

Subject indexing is a traditional task for libraries. It denotes the process of describing the contents of documents with appropriate concepts from a controlled vocabulary in accordance with certain criteria. It aims to cover the main topics exhaustively and describe them as precisely as possible, while seeking a condensed representation of the content that contains, for instance, roughly 5 to 8 concepts on average[6] [25, 20, 18, 11]. *Automatic subject indexing* attempts to implement this task algorithmically.

According to the Simple Knowledge Organization System (SKOS)[7], concepts represent abstract units of thought, and natural language expressions referring to concepts are called labels[8]. In this article, concepts of the controlled vocabulary will be referred to as descriptors[9]. SKOS vocabularies can provide additional information, for instance, links between concepts that encode hierarchical (broader/narrower) or associative (related-to) semantic relations.

The STW Thesaurus for Economics[4] is an example of such a controlled vocabulary in SKOS format. It is a wide-coverage bilingual resource (German and English) for economics, business studies and closely related subject areas. Version 9.02 of the STW[10] has more than 6,000 subject headings, more than 20,000 synonyms, and links broader, narrower, and semantically related concepts. Regarding broader and narrower concepts, the topology of the STW is a poly-hierarchy, that is, each descriptor can be linked to multiple broader descriptors. In addition, descriptors are categorized. They can be assigned to multiple subject groups (thsys), which are called categories in the remainder of this article. In contrast to descriptors, categories are linked with at most one broader category. Hence, the topology of subject groups is a mono-hierarchy.

4 Concept Drift

Concept drift has been studied in different contexts and there is a variety of terms in the literature for such phenomena. For clarification, we give a brief introduction to terminology and theory in the following subsection. Subsequently, we illustrate concept drift by analysing term-frequencies of documents at the German National Library of Economics in a practical setting.

4.1 Terminology

Concept drift has been formally defined for prediction tasks. This article primarily follows Gama et al. [8], but also borrows general terms which have been introduced for dataset shift [26]. In the following, let $\mathbf{x}$ be an input vector of features to predict the output $\mathbf{y}$. Let $\mathcal{D}^{\text{train}}$ be the training data, where correct values of $\mathbf{y}$ are known for each instance according to a specification of the task, and $\mathcal{D}^{\text{test}}$ data where the corresponding output vectors are assumed to be unknown. In this section, $\mathbf{x}$ may be interpreted as a term frequency vector of a document[11] and $\mathbf{y}$ as a vector that indicates which concepts belong to the document.

A basic principle behind typical applications of machine learning is the assumption that the training and test data sets have similar joint probability distributions, i.e.,

$$p_{\text{train}}(\mathbf{x}, \mathbf{y}) \approx p_{\text{test}}(\mathbf{x}, \mathbf{y}) \tag{1}$$

holds for the joint distributions of $\mathbf{x}$ and $\mathbf{y}$ on $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{test}}$, respectively. Concept drift, however, breaks

[6] The number of indexing terms depends on the particular content of a document and several other factors, such as individual institutional guidelines. As a consequence, averages reported in related work vary considerably. Some data sets are actually very similar to single-label document classification, as mentioned in Section 2.

[7] `www.w3.org/2004/02/skos`, accessed 10.11.2017

[8] In related work, especially in the domain of machine learning, the term "label" is often used for classes, which in turn represent concepts.

[9] This meaning of descriptors has been used in related work, but please note that descriptors denote special labels in SKOS.

[10] At the time of the experiments (Section 7), release 9.02 was the latest version. Version 9.04 of the STW has been released on June 21st, 2017.

[11] Different meanings of $\mathbf{x}$ will be used in other sections, for instance, in Section 5.

this assumption, and allows that the joint probability distribution of the training and test data set differ, i.e.,

$$p_{\text{train}}(\mathbf{x}, \mathbf{y}) \neq p_{\text{test}}(\mathbf{x}, \mathbf{y}) \tag{2}$$

which may be caused by hidden external factors. In contrast to Gama et al.'s concept drift definition [8], which emphasizes temporal aspects, the more general notion given above is closer to dataset shift [26], but this distinction is rather subtle.

Further categorizations of concept drift have been introduced, especially according to factorizations into conditional distributions $p(\mathbf{y}|\mathbf{x})$, $p(\mathbf{x}|\mathbf{y})$ and prior distributions $p(\mathbf{y})$, $p(\mathbf{x})$. *Real concept drift* refers to changes $p_{\text{train}}(\mathbf{y}|\mathbf{x}) \neq p_{\text{test}}(\mathbf{y}|\mathbf{x})$, i.e., conditional probability distributions, while *virtual drift* refers to changes in the covariates, i.e., $p_{\text{train}}(\mathbf{x}) \neq p_{\text{test}}(\mathbf{x})$, hence, we will prefer to use the term *covariate shift*. Notably, both phenomena may and often do appear in parallel [26].

In the context of subject indexing, related notions have been used by Tsoumakas et al. [35], who referred to "addition, deletion, merging of concepts" (explicit concept drift) and "altered semantics of concepts" (implicit concept drift), respectively. In this article, these terms should be interpreted as particularly linked to experimental settings. We use the term *explicit concept drift* for settings where documents concerning specific topics have been excluded from the training data. As a consequence, the concepts belonging to these topics are completely new in the test data. We will refer to settings where specific series or journals are excluded from the training data as settings with *implicit concept drift*. In such settings, different topics may be present as well as similar topics with different term distribution. Such settings may, however, also comprise data sets where the test documents are very similar to training documents. The intended primary effects of explicit and implicit concept drift settings therefore regard differences in prior distributions $p(\mathbf{y})$ and conditional distributions $p(\mathbf{y}|\mathbf{x})$ (real concept drift), respectively. Both types of concept drift are assumed to induce shifts in the distributions of covariates $\mathbf{x}$. Nevertheless, other side effects may be induced as well.

4.1.1 Visualization of Covariate Shift

In order to get an impression of concept drift between data sets, differences in their observed term frequencies, that is, covariate distributions $p_{\text{train}}(\mathbf{x})$ and $p_{\text{test}}(\mathbf{x})$, can be investigated. In this regard, terms, which are frequent in one corpus but infrequent in the other, are in the focus of interest, because they indicate concept drift. For finding such *characteristic terms* and revealing differences between corpora, a number of approaches have been proposed. Just to give a concrete example, Kessler [14] contributed a tool[12] which offers different options for term-weighting and scaling to create scatterplots. In particular, it includes a strategy that is based on the ranks of term frequencies.

In the remainder of this article, we utilize simple yet effective plots based on scaled term frequencies, which can be created similarly with the program provided by Kessler [14]. In the beginning, documents are sampled from both corpora. The contents (title and author keywords) are then tokenized and preprocessed, for example, changing title case to lower case. All further computation is based on the counts n_t^K of each term t in corpus $K \in \{A, B\}$, respectively. After scaling n_t^K to a virtual count m_t^K with respect to $T^* = 10,000$ tokens

$$m_t^K = n_t^K \cdot \frac{T^*}{T^K} \tag{3}$$

with the total number of tokens $T^K = \sum_t n_t^K$ in K, the position of the point representing term t is given by

$$x_t^K = \log(m_t^K + c) \tag{4}$$

with Laplace smoothing by c and $K \in \{A, B\}$ representing the x-axis and y-axis, respectively. Jitter was finally added to circumvent overlapping positions. Colors are assigned based on the difference $\Delta_t = x_t^A - x_t^B$ and alpha values for color are derived from the distance to the origin of the coordinate system.

As a consequence, the number and degree to which terms are plotted away from the diagonal can be interpreted as a measure of concept drift. For instance, in Figure 1, the terms "loyalty" and "brand" are relevant to capture the contents of the test documents, however, they occurred rarely in the training documents. By contrast, the frequencies of function words like "the" and "and" remain almost stable.

A certain degree of difference in distributions must be considered as expected noise. Because term frequencies are typically distributed by a power law (Zipf's law), many terms are infrequent, hence, new terms are a natural effect of randomly sampling training and test documents. For this reason, our experiments in Section 7 will compare concept drift settings to random sampling settings.

4.2 Concept Drift in Subject Indexing

At digital libraries, subject indexing datasets where training data and test data differ may occur for different reasons, such as

[12] `https://github.com/JasonKessler/scattertext`, accessed 24.08.2017

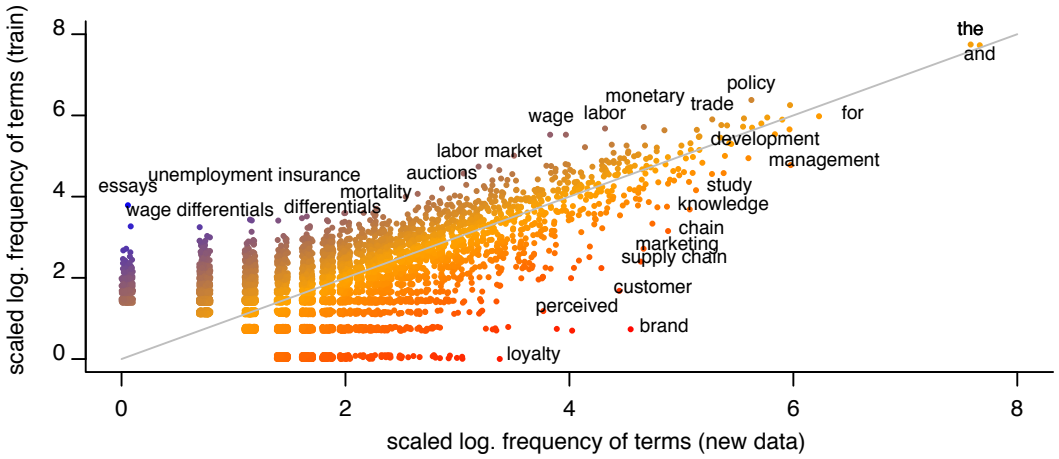


Fig. 1: Visualizing drift of covariates. Terms in titles and author keywords in practice. Professionally indexed documents (training, y-axis) vs. documents not yet indexed with STW descriptors (new data, x-axis).

1. externally caused temporal drift,
2. latent sample selection bias, or
3. revised outcome specifications.

Changes over time (*temporal drift*) are known properties of publication practice. For instance, temporal trends in economics publications have been studied by Kosnik [15]. According to her results, publications in economics have increasingly dealt with mathematical methods. Therefore, shifts in research attention involve even high-levels of abstraction, in this case, high-level categories of the JEL classification system[13]. In addition to varying interest in research topics, language evolves over time, which comprises changes in word meanings, their surface forms, and syntax. Such changes have been studied, for instance, in the context of digital humanities [12, 30], where differences can be detected and tracked over long time spans. While these phenomena obviously can affect subject indexing, they are not in the focus of this article. Contrary to digital humanities, we consider data from shorter spans of time, that is, decades rather than centuries, here.

Sample selection bias can be caused, for instance, by indexing preferences. Libraries may be specialized to certain subjects or have indexing preferences regarding particular topics, journals, geographical regions, time spans, authors, or genres, just to name a few. Such an institutional focus can influence the selection of documents that are indexed by humans, hence, potentially introducing a bias on priors $p(\mathbf{y})$ against the library's complete catalog. Assumptions on independent and identically distributed data can therefore be violated.

In addition, *revised outcome specifications*, such as altered indexing rules and guidelines, or controlled vocabulary changes (addition, removal, alteration of concepts) are actions that consciously control how documents with the exact same words are indexed. This certainly implies modifications of $p(\mathbf{y}|\mathbf{x})$ (real concept drift). Different from the latent externally induced temporal drift that was mentioned before, these shifts are caused deliberately, thus, related change events may be reconstructed and regarded for historic data, for instance, by creating subsets of documents by date according to releases of the controlled vocabulary and indexing rules, and training different classifiers accordingly. This concept drift adaptation approach decreases, however, the number of training examples for each descriptor. Since typically many descriptors are rare, this type of drift may be completely neglected instead, in favor of larger sets of training documents. In the extreme case, at the moment when a new version of the controlled vocabulary or new guidelines are released, no corresponding training examples are available. In this case it may be reasonable to assume that for most of the descriptors, the data of the outdated guidelines will be a sufficient substitute until more appropriate data has been produced by professional human indexers. Although optimal adaptation cannot be reached with this strategy, it may, however, be an appropriate interim solution. Further investigation of this topic will be subject of future work.

Furthermore, it should be noted that the structure of the controlled vocabulary and knowledge representation in general can have substantial influence on the appearance of concept drift. For instance, while it is difficult to make clear distinctions between named entities, concepts, and even theories or genres, their aspects appear in controlled vocabularies. Similar to language in general, where closed classes of part-of-speech are

[13] Journal of Economic Literature (JEL) codes: `https://www.aeaweb.org/econlit/jelCodes.php`, accessed 10.11.2017

opposed to open classes, some parts of thesauri may change more rapidly than others.

Since temporal drift is an inherent and thus timeless aspect of research, publishing and its indexing, we argue that effects of concept drift should gain more attention.

Covariate Drift in Practice

By example, we now turn to data of the German National Library of Economics (ZBW) and a special subset of documents where keywords are available which have not been specified by a known controlled vocabulary. Some of these documents in the catalog have been indexed additionally by professional staff, hence, they may be used as training data. The other documents will be named new data here. In this study, we focus only on documents with meta-data in English.

Figure 1 depicts the term frequencies in the training data versus the new data using the visualization described above. As can be seen, the more frequent a term occurs in a data set, the more likely it is that it also appears in the other data set. Nevertheless, certain terms having a meaning relevant to subject indexing, like "supply chain" or "brand", seem to be rare in the training data, but frequent in the new data. We will return to this plot with a possible interpretation in Section 7.

5 Analysis of Indexing Systems

This section analyses architectures of indexing systems and outlines strengths and weaknesses that can be derived independently of specific implementations. It focuses on the way background knowledge is used and how the approaches scale with respect to growth of the controlled vocabulary. We will base our discussion on the aspects depicted in Table 1, namely (A1) the amount of training data required (low is better), (A2) whether previously unseen concepts can be predicted (desirable), (A3) whether synonyms can be predicted (desirable), (A4) whether ambiguity can be resolved (desirable), (A5) whether relations of concepts in the thesaurus are used (desirable), and (A6) the applicability for short texts. While (A1), (A2) and (A3) will be discussed first for each type of approach, (A4), (A5) and (A6) will be discussed separately in successive paragraphs.

For the discussion we will use the following, small example of a document with author keywords and professional indexing terms:

Title: Analysis of the German gas price from 1970 to 1980. Author keywords: Germany ; energy pricing ; gas ; 70s. Indexing terms: c:gas price ; c:Germany.

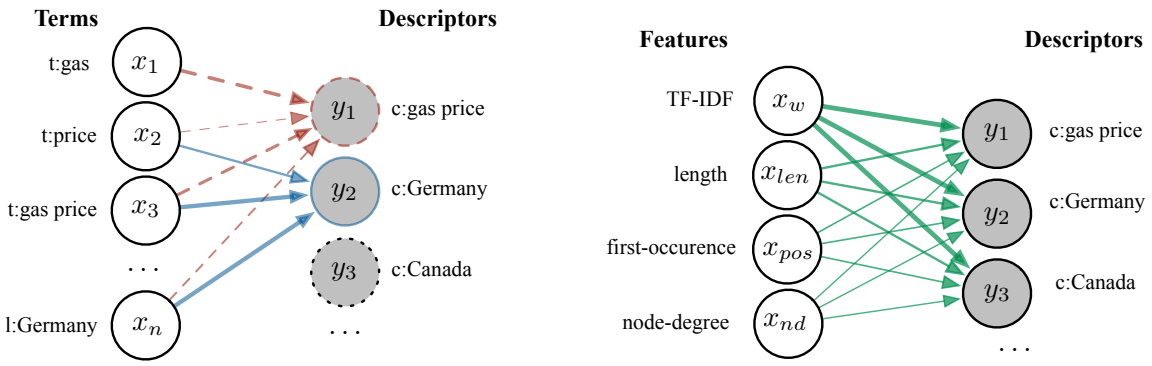
Table 1: Pros (+: advantage) and cons (-: disadvantage) of lexical (L) and associative (A) system architectures according to challenges in automatic subject indexing. (Copyright © IEEE, see footnote 5)

	Aspect	L	A
A1	Amount of required training data	++	-
A2	Prediction of unseen concepts	++	--
A3	Prediction of synonyms	--	++
A4	Ambiguity	o	+
A5	Exploitation of thesaurus relations	+	o
A6	Applicability to short texts	o	o

Different prefixes are used to refer to different types of features: terms/word n-grams (t), dictionary matches to labels of concepts (l), concepts, i.e., descriptors (c).

Figure 2a shows a prototypical **associative indexing** system for the example document. On the left, we can see features like the term feature "t:gas" or a match of a certain concept label "Germany" that encode the document. Typically, one feature is created for each unique n-gram of the training documents resulting in a large number of features. On the right hand side are class nodes that encode concepts that might be assigned by the system, for instance, "c:gas price". Under this representation of documents and their concepts, systems operate on sparse representations, that is, most entries of the corresponding document-feature matrices are zero. A variety of machine learning algorithms may be used to determine how features and individual concepts relate to each other. Generally, co-occurrence statistics are used to describe concepts and discriminate them from other concepts. In this methodology, it is therefore possible to derive associations from data, such as that the term "t:FRG" is a positive indicator for the descriptor "c:Germany (Federal Republic)". Broadly speaking, unknown synonym expressions can be learned from the data (A3). Based on the commonly used binary relevance approach[14] [29, 10], parameters that finally determine if a concept is assigned are learned independently for each descriptor. In Figure 2a, parameters of a classifier (encoded by color) and their weights (encoded by line thickness) are shown as arrows between terms (nodes x_i) and descriptors (nodes y_i). No weights have been learned for y_3 (c:Canada) because no training instance was available for this concept. As a consequence, this concept can not be assigned to any document (A2). Even if we add concept features for matches against the thesaurus to the feature vector [17, 11] to encode background knowledge, descriptor-specific parameter learning makes it impos-

[14] Links to approaches that relax this constraint are given in the related work, see Section 2.



(a) Associative indexing

(b) Lexical indexing

Fig. 2: Comparison of architectures by example. a) In associative indexing, the learning algorithm learns relations between features, which are terms (t:) or dictionary matches (l:), and descriptors (c:) for each descriptor independently. b) In lexical indexing, features are computed for concept candidates derived from the document's terms. Feature weights are shared among all descriptors for classification. (Copyright © IEEE, see footnote 5)

sible to assign concept "c:Canada" when no training example is available for this descriptor. For each descriptor in the thesaurus, at least one training example is required (A1). In fact, reliable estimates typically demand more data.

A prototypical **lexical indexing** system is illustrated in Figure 2b using features from KEA++ [20] as an example. Based on lexical knowledge from a thesaurus, the system first extracts several concept candidates (c:gas price, c:Germany, c:Canada) from the text by applying dictionary matching. Feature values $(x_{\mathrm{w}}, x_{\mathrm{len}}, x_{\mathrm{pos}}, x_{\mathrm{nd}})$ are then computed for each candidate, and decisions on the output are finally made by repeated application of the same classifier, as shown by duplicates (y_1, y_2, y_3) of the same node template for all concept candidates in Figure 2b on the right hand side. Just for illustration, let us consider classification based on the computation of real-valued scores by linear combinations $y_i = w_1 \cdot x_{\text{tf-idf},i} + w_2 \cdot x_{\text{len},i} + w_3 \cdot x_{\text{pos},i} + w_4 \cdot x_{\text{nd},i}$ with weights $w_1 = 2, w_2 = 1.2, w_3 = 0.7, w_4 = 0.34$ (as an example). The final descriptor assignment is then based on this score and a threshold τ, such that $y_i > \tau$ triggers the assignment of the ith concept. Please note that the weights and the threshold are the same for all instantiations of the template. Put in different words, the lexical system shares the same feature weights (green arrows) for all possible descriptors. As a consequence, the system learns weights that are re-usable, even for previously unseen concepts, like "c:Canada" in the example. This fact is one of the main differences to associative approaches. Consider that we apply the system to a new document that contains the term "Canada" which is recognized during concept candidate generation by dictionary matching. The system then computes TF-IDF, length, first-occurrence and node-degree features for this match. Subsequently, the same parameters that have been optimized for other descriptors are utilized to decide if the descriptor of Canada should be assigned. It can successfully be added to the output list of descriptors (A2). As can be seen, there is only a small number of features, in this special case four, thus only a limited number of parameters have to be fit. Furthermore, the feature representation will be rather dense because the four feature functions in the example will often have non-vanishing values for candidates. For these reasons, the conditions for reliable parameter estimation are good. Only a few documents are required for training [20] (A1). But it comes at a cost. The approach is unable to learn synonymous expressions from data (A3). It is completely built upon and restricted to the dictionary matches against the controlled vocabulary.

Directly compared to each other regarding aspect (A1), the associative system is supposed to scale at least linearly in the number of required training examples when the controlled vocabulary size is increased while for the lexical systems this remains constant.

Natural language is inherently ambiguous (A4) and word senses have to be determined in order to understand a text. Associative approaches can learn to solve this task using arbitrary words in context, but remain limited to known concepts and words from training data. Lexical approaches depend in their performance on the controlled vocabulary. If enough candidates can be extracted, features like node-degree or descriptor co-occurrence expectations may enable to determine the correct sense of a phrase.

To underline further differences, let us consider the use of relations between concepts retrieved from background knowledge (A5), like "c:price" is broader than "c:gas price". As shown in Figure 2b, it has been proposed by Medelyan et al. [20] to compute a *node-degree* feature that measures how strong a concept candidate is connected to other candidates in the same document. Parameters are shared among descriptors and learning is therefore based on many examples. The importance of this feature can be confidently estimated and generally applied. In associative systems, concept features can be activated based on different schemes [17, 11]. Learning and prediction remain, however, restricted if only concepts from the training data can be predicted like in kNN classification or if individual classifiers are learned for each descriptor.

In principle, both associative and lexical approaches can be applied to short texts (A6), however, certain phenomena might be more pronounced and should be considered during configuration when only a few terms are available. For instance, the node-degree feature of lexical systems may not find enough related candidates in very short text for meaningful operation.

In summary, we conclude, that lexical classification and associative classification provide distinct capabilities in order to achieve accuracy and scalability. A comprehensive overview of advantages and disadvantages of both systems can be found in Table 1.

6 Fusion Architectures

In the last section, we have seen that approaches that are solely lexical or solely associative fail on some challenges of automatic indexing but also have individual strengths. Therefore it seems reasonable to attempt a fusion of both approaches by combining the individual predictions. The interesting questions are, however, how fusion is actually realized and which pitfalls have to be avoided.

The top level simplified design of the proposed fusion architecture is depicted in Figure 3. First, different candidate sets are produced: by an associative component (center, left) that leverages a large set of professionally indexed documents, and by a lexical system (center, right) that relies on background knowledge from a thesaurus. Then, the fusion layer (below) is responsible for combining these predictions. The most interesting property of this layer is the *descriptor-invariant decision function* [32], i.e., a function that allows to perform predictions for all (also unseen) descriptors. Optionally, the fusion module may additionally consult the knowledge base or the professionally indexed documents for its decisions and use a descriptor-specific fusion component.

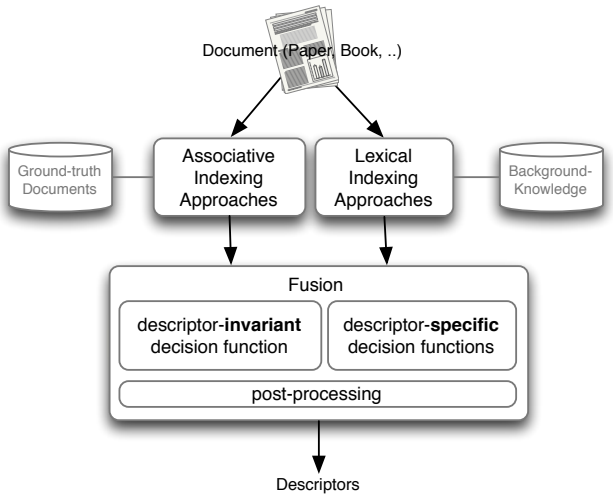


Fig. 3: Generic schema of a fusion system.

Within the fusion layer, it is crucial how the predictions are combined. On the one hand, one may learn on a basis of descriptors (descriptor-specific fusion), for example, learning mapping tables [13] using confidence tests. In a similar but different manner, we can simply compute for each descriptor c and method m the support (number of documents with c assigned by m) and confidence (number of c correctly proposed by m divided by its support) for each descriptor c based on held-out data of the training set. Descriptors that surpass a minimum support and a minimum confidence may then be added by m to the final output in a production setting (testing). This simple strategy, in the following referred to as *Rhack*, is slightly different from mapping tables that map descriptors to methods [13]. While the latter may learn that the concept "theory" is better predicted by the associative component than by the lexical component and therefore will choose to *always* handle it by the associative system, Rhack will simply join their predictions and assume that both are reliable. We suggest that both kinds of behavior are not optimal in general because they are again restricted to the set of known descriptors from the training documents. They will not be able to determine a suitable predictor for the term "Canada" if this term is not present in the training documents. Even if dictionary matching is used per default (cf. [13]), mapping tables can leave benefits of complementarity aside because, depending on the actual implementation, only one single method is chosen per concept.

Therefore, a fusion decision function should be implemented that is invariant to descriptors. In order to

investigate the potential of the proposed design, we construct a very straight-forward system. We study the *union* of predictions per document. This strategy is derived from the idea of setting the above-mentioned minimum confidence and support to zero in the fusion layer, but expands predictions to previously unknown concepts. Each subsystem may, however, still filter its predictions by an individual confidence threshold. This is indeed essential to guarantee high precision in the fusion system. The union approach is straightforward, however, it has some interesting aspects and especially enables us to explore if higher recall can be reached by fusion. Following the discussion of existing architectures in the previous section, we observe that:

- Associative systems may suffer from low recall, because the data they learn from is likely to be insufficient. Terms and concepts follow power laws, hence, many concepts and terms are infrequent.
- Lexical systems may suffer from low recall, when the knowledge base lacks synonymous expressions, especially when texts are short and therefore less candidates are generated per document.

For these reasons, gaining recall in the fusion layer seems to be crucial and it may be a promising way to join predictions for better overall performance.

Besides choosing between concept candidates from the subsystems, we also investigate *post-processing* aspects of the fusion layer. During fusion, systematic errors of individual modules might be corrected with supervision that builds upon predictions, professionally indexed documents, and background-knowledge from the thesaurus. Inspired by ideas from transformation-based error-driven learning [4], we investigate a *transformation-rule learning* module. For each pair of categories (k_1, k_2) in the thesaurus, it counts cases on held-out data of the training examples where a descriptor $c_1 \in k_1$ was predicted erroneously while a related descriptor $c_2 \in k_2$ was missed at the same time. It then attempts to increase performance on the training data with a transformation rule (switch every prediction of c_1 by c_2). If it succeeds, this rule is added to a list of rules that are used in production to index new documents. For instance, we may learn a rule that replaces candidates c_1 by c_2 if c_1 is a geographic adjective or language (e.g. "German"), c_1 and c_2 are related concepts as defined in the thesaurus, and c_2 is a geographic name (e.g. "Germany"). Interestingly, such transformation rules may predict previously unseen concepts when they consider types of descriptors instead of descriptors themselves; the example rule above applies to "Canadian" even when "Canada" was not part of the training data.

6.1 Concept Drift and Fusion

Before we turn to specific implementations of the fusion framework, we would like to make a note on the importance of fusion with respect to concept drift.

Recap Section 4, concept drift may in particular comprise shifts in priors that may lead to a number of vanishing and emerging descriptors, as well as differences in co-occurrence statistics. Therefore, we hypothesize that aspects (A1) and (A2) in Table 1 will be particularly relevant under concept drift. As a consequence, lexical approaches are supposed to handle shift in priors, for example, caused by sample selection bias, better than associative approaches because of descriptor-invariance of lexical systems. Similarly, we expect that fusion systems are more robust to concept drift than associative systems. The impact of these relations is, however, determined by several environmental factors. For these reasons, experiments are conducted and described in the remainder of this article.

6.2 Implementation

In the presented framework, associative predictions and lexically-based prediction modules may be implemented by different methods. In the following experiments, we especially consider two state-of-the-art approaches for each type: maui [21] to produce predictions with lexical background knowledge and approaches related to SGD-SVM [36] for prediction in an associative way.

Inside of the lexical layer, maui [21] provides a mature thesaurus-based system with a rich set of features that goes beyond simple dictionary matching. In our case, it can, however, be assumed that different features are required to realize the full potential of short texts like titles or author keywords. For instance, maui's span feature aims to weight terms higher that are mentioned in the abstract and the conclusions, which are however not accessible in this case. We leave the invention and integration of new features for future work and suspect that maui's supervised learning method (bagged decision trees [3]) will still be able to create a robust prediction component, even when applied to short texts.

7 Experiments

With the experiments we wanted to answer the following three experimental questions:

i) How do fusion systems compare to associative and lexical approaches in terms of overall accuracy?
ii) To which extent are the approaches robust to explicit concept drift?

iii) To which extent are the approaches robust to implicit concept drift?

Explicit concept drift is modelled by a test data set containing descriptors from specific categories that are not present in the training data set. To assess implicit concept drift we evaluate the trained models on an unknown series of documents, which may cover different topics. We perform the experiments on short texts from the economics domain and using the STW thesaurus (cf. Section 3).

7.1 Data Set

Our data set consists of documents represented by their titles and author keywords only. This information is available even in indexing scenarios where abstracts or full-texts are either missing (in the case of books) or not accessible due to legal aspects. We represent the documents as described in Section 5. The complete sample contains 20,195 documents, indexed by professional indexers. Indexers assigned 5.85 ($SD = 1.84$) descriptors per document on average. 94% (19,054) of the documents have a unique combination of descriptors assigned.

To compare i) the overall performance of the different approaches we split the data set randomly into training and test sets using 5-fold cross-validation (data set denoted by $\mathcal{D}_{\text{shuffle}}$).

In order to measure the influence of ii) *explicit concept drift*, we created data sets denoted $\mathcal{D}_{\text{cat}}$, where we split the documents according to certain subthesauri (categories), that is, subject fields. We used sets of classification scheme codes ("thsys" codes) of the STW for which we ensured that they were not used during training. For instance, one training set of $\mathcal{D}_{\text{cat}}$ does not contain documents with descriptors from the field "marketing" (thsys: B.07), but all test documents cover some descriptors from this category, for instance, market share, competition, or customers. Consequently, this setting emphasizes the zero-shot learning task.

To investigate the influence of iii) *implicit concept drift*, we split documents into sets $\mathcal{D}_{\text{series}}$ according to publication series. For example, one single working paper series which covers subjects like "regional business growth programmes" or "human capital" is omitted from training. The corresponding test set includes only documents from this series.

Table 2 provides an overview of the different data sets. The average number of assigned concepts is the same on training data and testing data for the random splits $\mathcal{D}_{\text{shuffle}}$, but it differs on $\mathcal{D}_{\text{cat}}$ and $\mathcal{D}_{\text{series}}$, respectively. The explicit and implicit concept drift settings have larger training sets on average, but the size of the corresponding test sets varies. For instance, the test subsets of $\mathcal{D}_{\text{series}}{}^{(\text{test})}$ contain 4742, 748, 415, 385, and 385 documents.

Table 2: Properties of settings with respect to professional indexing. $|\{\mathcal{D}_i\}|$: number of partitions (folds). $|\bar{\mathcal{D}}|$: average number of documents. $|\bar{\mathcal{L}}|$: average number of unique descriptors. $|\bar{\mathcal{Y}}|$: average number of descriptors per document. (Copyright © IEEE, see footnote 5)

Setting	$\lvert\{\mathcal{D}_i\}\rvert$	$\lvert\bar{\mathcal{D}}\rvert$	$\lvert\bar{\mathcal{L}}\rvert$	$\lvert\bar{\mathcal{Y}}\rvert$
$\mathcal{D}_{\text{shuffle}}{}^{(\text{train})}$	5	16,156	3,848.8	5.85
$\mathcal{D}_{\text{shuffle}}{}^{(\text{test})}$	5	4,039	2,777.2	5.85
$\mathcal{D}_{\text{cat}}{}^{(\text{train})}$	5	17,490	3,812.8	5.78
$\mathcal{D}_{\text{cat}}{}^{(\text{test})}$	5	2,705	1,946.0	6.26
$\mathcal{D}_{\text{series}}{}^{(\text{train})}$	5	18,860	3,950.0	5.82
$\mathcal{D}_{\text{series}}{}^{(\text{test})}$	5	1,335	1,205.4	6.54

Concept Drift

In order to get an impression of concept drift (none, explicit, implicit) in the data, Figure 4 depicts term frequency distributions as described in Section 4.1.1. For each corresponding setting (shuffle, cat, series), we created one diagram based on one partitioning, with a maximum of 5000 randomly sampled documents per partition. We set the minimum term frequency to $n = 5$ and the Laplace smoothing to $c = 1$. Sentence boundary characters were removed. Tokens were converted to lowercase, if they were title-cased and had at least two characters.

As can be seen, the shapes of the diagrams differ considerably. Following our expectations, plot b), which refers to explicit concept drift, has more characteristic terms than the shuffle setting (no concept drift), shown in a). Interestingly, this also holds for the comparison between the folds shown in c) (implicit concept drift) and a), hence, concept drift on new series may be similar to explicit concept drift in particular settings.

Comparing all three of these plots visually with Figure 1, it seems that the term frequencies of the practical setting, shown in Figure 1, are spread away from the diagonal more similar to explicit concept drift 4b) and implicit concept drift 4c) than to the shuffle setting 4a). This indicates that the drift in the practical setting is irregular and unlikely to be noise, hence, underlining the relevance of this study.

Opposed to Section 4.2, availability of suitable descriptors is given for all folds. Consequently, we can depict changes between prior probabilities $p_{\text{train}}(\mathbf{y})$ and $p_{\text{test}}(\mathbf{y})$ in a similar way, as shown in Figure 5. For these plots, we set the minimum concept frequency to

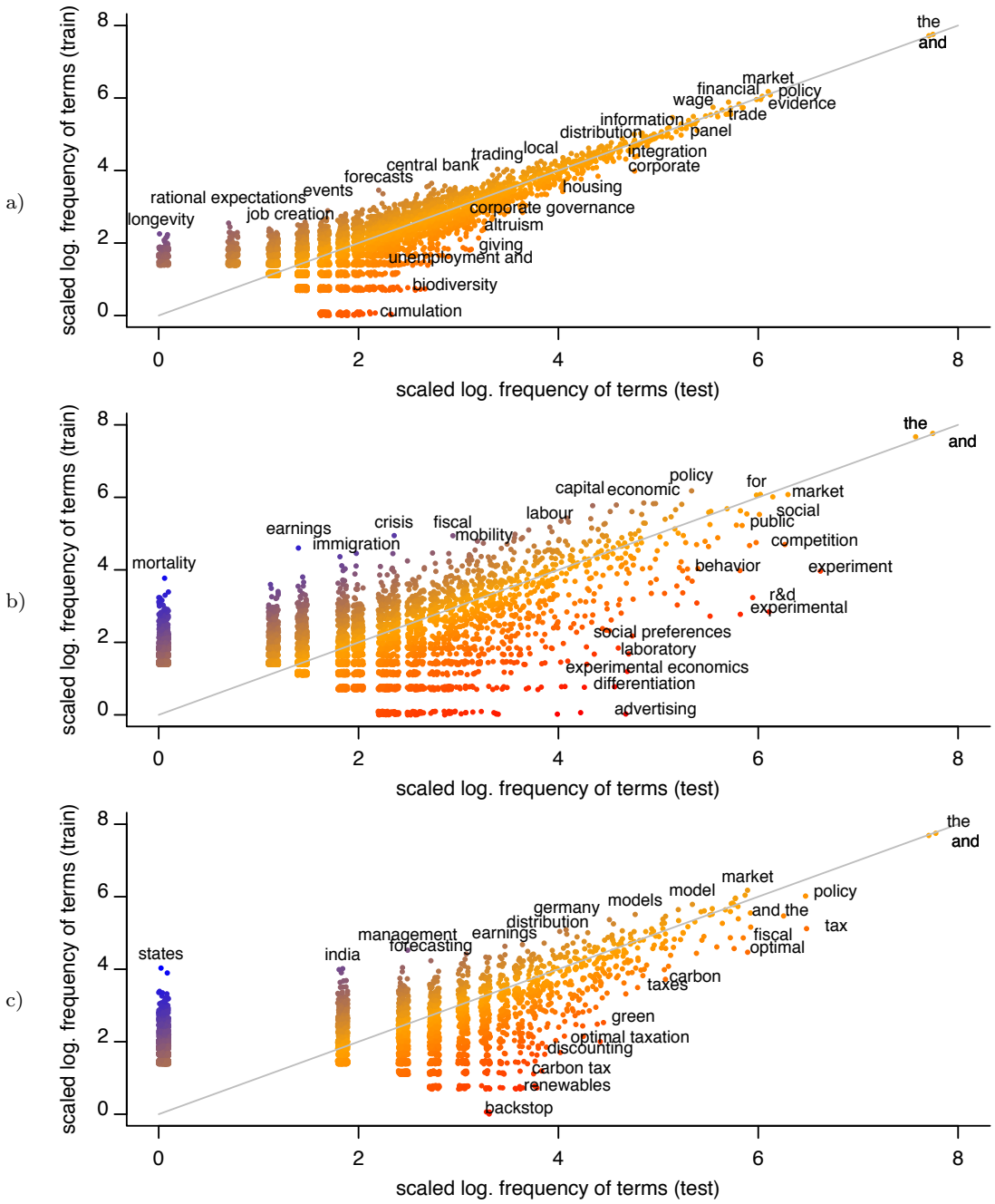


Fig. 4: Visualisation of concept drift. a) Random data set splits (shuffle), no concept drift, example data sets $\mathcal{D}_{\text{shuffle}}^{(\text{train},1)}$ vs. $\mathcal{D}_{\text{shuffle}}^{(\text{test},1)}$. b) Explicit concept drift on data sets $\mathcal{D}_{\text{cat}}^{(\text{train},1)}$ vs. $\mathcal{D}_{\text{cat}}^{(\text{test},1)}$, c) Implicit concept drift on data sets $\mathcal{D}_{\text{series}}^{(\text{train},1)}$ vs. $\mathcal{D}_{\text{series}}^{(\text{test},1)}$.

$n = 1$. It can be seen that the concept drift settings b) to d) differ considerably from the shuffle setting a). In particular, the explicit concept drift setting shown in b) poses a hard challenge with descriptors that are missing in the training data. Nevertheless, also implicitly induced concept drift has clearly shifted the concept prior distributions in the shown data sets c) and d).

7.2 Evaluation Metrics

We use common metrics [29] which can be computed in total (micro-average), per concept (macro-average), or per document (sample-based average): precision (correctly predicted descriptor assignments divided by all predicted descriptor assignments), recall (correctly predicted descriptor assignments divided by all descriptors

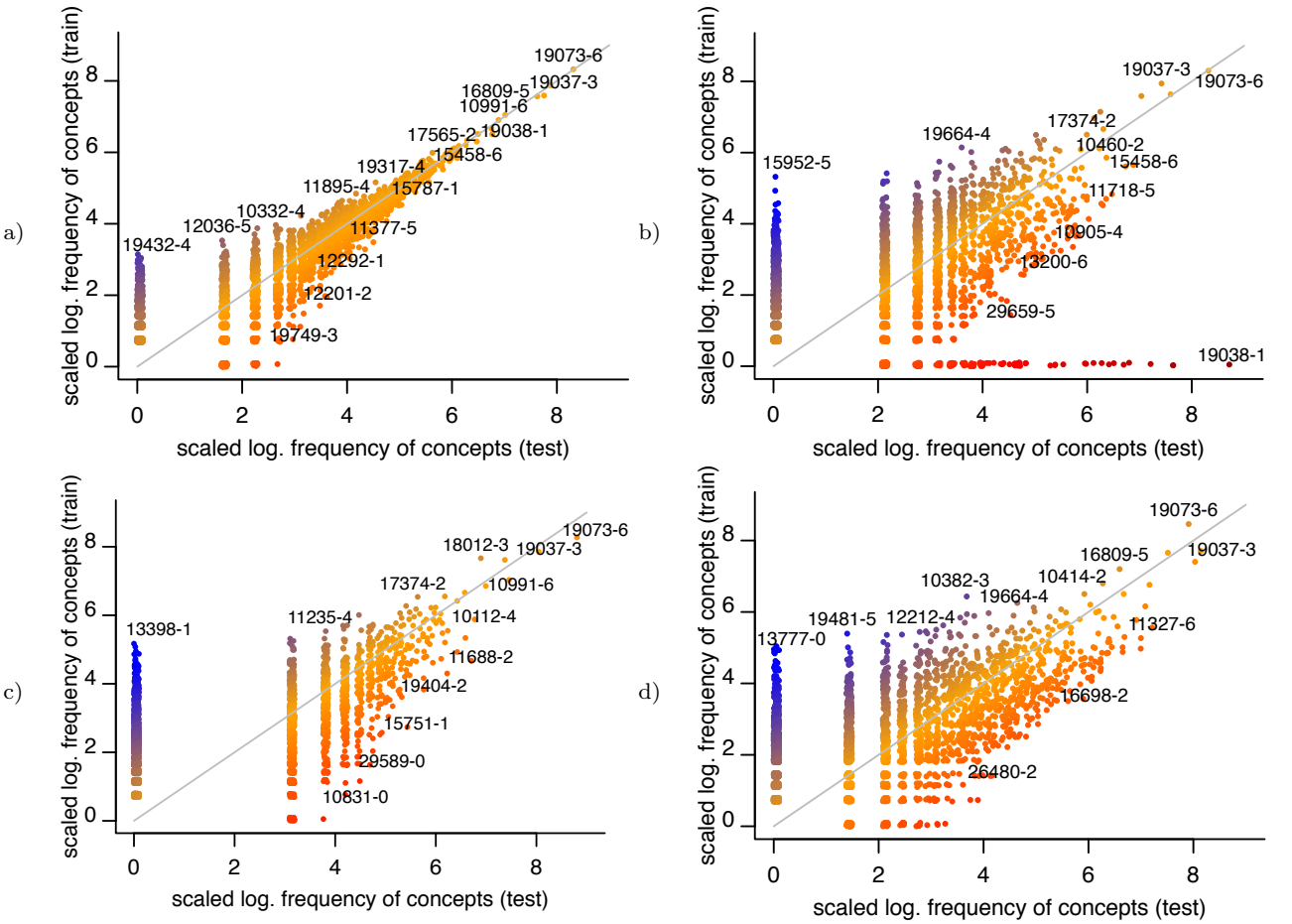


Fig. 5: Visualisation of shift in priors of concepts: a) Random data set splits (shuffle), no concept drift, example data sets $\mathcal{D}_{\text{shuffle}}^{(\text{train},1)}$ vs. $\mathcal{D}_{\text{shuffle}}^{(\text{test},1)}$. b) Explicit concept drift on data sets $\mathcal{D}_{\text{cat}}^{(\text{train},1)}$ vs. $\mathcal{D}_{\text{cat}}^{(\text{test},1)}$, c) Implicit concept drift on data sets $\mathcal{D}_{\text{series}}^{(\text{train},1)}$ vs. $\mathcal{D}_{\text{series}}^{(\text{test},1)}$. d) Implicit concept drift on data sets $\mathcal{D}_{\text{series}}^{(\text{train},3)}$ vs. $\mathcal{D}_{\text{series}}^{(\text{test},3)}$. For instance, concept 10831-0 ("tax haven", see `zbw.eu/stw/descriptor/10831-0`) occurred more frequently in the test set $\mathcal{D}_{\text{series}}^{(\text{test},1)}$ than in the training set $\mathcal{D}_{\text{series}}^{(\text{train},1)}$.

assigned by professional indexers), F_1 score (harmonic mean of precision and recall). Since macro-averaging metrics are not weighted by concept counts, they show if concepts are recognized accurately independently of their frequencies in the test sets.

Whether precision is more relevant than recall or vice versa depends on individual application requirements. For this reason, we provide details regarding both metrics. For the sake of simplicity, F_1 values are employed to summarize the results, considering precision and recall as being equally important.

7.3 Configurations

As two basic **lexical systems**, we implemented dictionary matching approaches: a simple matching algorithm that only considers phrases between stop words, denoted DICT, and MONQ which accesses a dictionary matching library[15] that considers morphological variants of terms and which was used in related work [13]. As a strong lexical baseline, we chose MAUI[16] [21]. The maximum number of concepts to assign was set to $k = 15$ and the minimum confidence was set to $c = 0.1$. Please note, however, that MAUI is typically applied to full text rather than short text.

Associative systems were realized by binary relevance (BR) approaches. We chose to use BRLR (logistic regression classifier) and BRSVM (support vector machines) trained by stochastic gradient descent (cf. Wilbur et al. [36]). Both, BRLR and BRSVM, were

[15] `https://github.com/HaraldKi/monqjfa`, accessed 10.11.2017

[16] `https://github.com/zelandiya/maui`, accessed 10.11.2017

configured with word n-gram features between stopwords.

RHACK (cf. Section 6) is a meta-learning approach which is similar in mind to [13]. We configured it to enrich predictions made by BRLR with the dictionary matching of DICT, adding only confident DICT predictions to the list of descriptors created by BRLR. On the training data, it therefore determines all concepts with minimum support (min.sup = 20) and minimum confidence (min.conf = 50%). These estimates for DICT predictions per concept rely on training data and implicitly measure a degree of association between terms and descriptors. As a consequence, it belongs to the associative system architectures.

Fusion approaches combining lexical and associative characteristics have been realized by combining the predictions of BRLR and DICT (short form: D) as well as of BRLR and BRSVM with MAUI using the *union* strategy described in Section 6. The names of these fusion systems are given by BRLR+D, BRLR+MAUI, and BRSVM+MAUI, respectively.

For DICT, BRLR+MAUI and BRSVM+MAUI, we additionally applied the **transformation** described in Section 6 which led to systems in the following denoted by the suffix T or *transform*. Due to the runtime of the quickly realized implementation of transformation rule learning[17], transformations were only determined based on the DICT method on the first fold and restricted to high-level categories of the thesaurus. Because the number of examples per category is expected to be high, we suspected that these rules are representative for all data sets and settings.

For the experiments we used python and the scikit-learn library[18] which support BRLR and BRSVM. For RHACK, we additionally used a script written for the R statistics package. MAUI and MONQ were applied with Java.

7.4 Results

Figure 6 compares distributions of the number of predicted concepts per document for each setting. For the purpose of illustration, only one method (BRLR, MAUI, BRLR+MAUI.T) is shown for each type of architecture (L, A, F). It can be seen, that all automatic methods (L, A, F) predicted less concepts than human indexers (truth). Especially the binary relevance approach (A) produces a considerable amount of documents which only consist of a few concepts. Fusion of predictions (F) lead to more human-like indexing in terms of the plain number of concepts per document.

Table 3 lists the results for all data sets and approaches, supplemented by Figure 7 which focuses on sample-based averages and gives a visual impression of how systems perform, with the focus on BRLR, MAUI, BRLR+MAUI, and BRLR+MAUI.T for the sake of readability.

Best values are marked bold in the table, showing that fusion approaches (arch.: F) that combine binary relevance approaches and MAUI were superior to lexical (arch.: L) and associative approaches (arch.: A) on all settings in terms of sample-based F_1 score and concept-based F_1 score. In almost all cases[19] this difference is statistically significant (paired t-test to the best performing algorithm), as indicated by arrows in the table. Across all settings, associative approaches (binary relevance methods and RHACK) achieved often significantly higher precision than other methods, however, they only predicted less than 3 descriptors per document on average. Recall of fusion systems outperformed associative as well as lexical approaches. These differences can also be recognized in Figure 7.

When training and test data were selected to reflect explicit concept drift (experimental question ii), the associative systems deteriorated considerably while MAUI was more stable (compare Figure 7). To highlight specific details, Figure 8 depicts results of two explicit concept drift settings, that is, category G.01 (Europe)[20] on the top and B.07 (marketing)[21] on the bottom, where evaluation has been constrained to concepts belonging to these specific categories only (left: B.07, right: G.01). Consequently, zero-shot learning settings can be found in the panels at the top-right and the bottom-left. For the sake of clarity, only four characteristic methods (listed on the left) have been regarded. Notably, it can be seen that (1) F_1 measure of the associative approach (A: BRLR) vanished for the zero-shot learning tasks, (2) the fusion approaches (prefix "F", combinations of A and L) improved the performance, and (3) modifications by transformation rules lead to improvements under special circumstances, for instance, with regard to category G.01 (Europe) and the zero-shot learning setting (top right panel).

Concerning experimental question iii), the implicit concept drift setting showed results that are similar to $\mathcal{D}_{\text{shuffle}}$, however, they seem to be more diverse, as can be seen in Figure 7.

[17] several hours on several thousand documents

[18] `www.scikit-learn.org`, accessed 10.11.2017

[19] In some cases, the data was not shown to be normally distributed (Shapiro-Wilk test, $p < 0.05$), thus the assumptions for t-tests were not met.

[20] `http://zbw.eu/stw/thsys/70002`, accessed 10.11.2017

[21] `http://zbw.eu/stw/thsys/70041`, accessed 10.11.2017

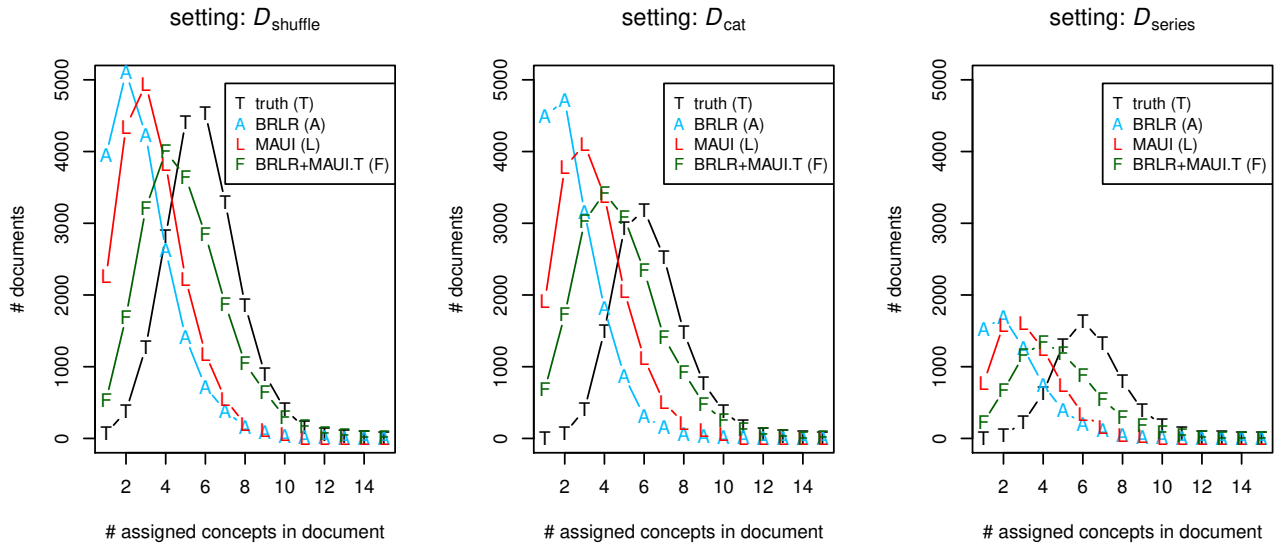


Fig. 6: Comparisons of distributions regarding the number of assigned concepts per document for random data set splits (left), explicit concept drift (center), and implicit concept drift (right).

As mentioned in Section 4.2, we expected that the type of descriptors may have impact on system performance. Therefore, we looked into detailed aspects of predictions for some folds. We observed that particularly concepts like "theory" or "estimation", which are amongst the most frequently assigned descriptors by human indexers, and which are rarely mentioned literally in the title of documents, have a very poor performance according to approaches that are based on dictionary matching. Detection of these concepts is especially improved by statistical approaches. If descriptors are infrequent but used literally and unambiguously by authors in the title, lexical methods outperformed statistical approaches. An example of such a descriptor was "elasticity of substitution".

7.5 Discussion

Considering the questions i)-iii) posed in the beginning of Section 7, results showed the following:

The proposed descriptor-invariant fusion is i) superior to the associative and lexical systems in terms of F_1, which is foremost attributed to changes in recall. The union of individually proposed descriptors per document substantially increased the overall recall. Hence, concepts proposed by the systems are at least partly non-overlapping. With the union strategy, the average number of assigned descriptors comes closer to how professional indexers act. Secondly, the union also retains high precision assignments, especially from the associative component.

With regard to question ii) and iii), fusion makes predictions more robust against concept drift as can be seen in Figure 7, supported by the details highlighted in Figure 8. Fusion is backed up by MAUI [21], which seems to be a robust choice for the lexical component of the system. Implicit and explicit concept drift were handled with lower variance by MAUI ($F_1 \approx 0.3$ on $\mathcal{D}_{\text{shuffle}}$, $\mathcal{D}_{\text{cat}}$, $\mathcal{D}_{\text{series}}$) compared to associative systems. The category setting $\mathcal{D}_{\text{cat}}$ (explicit concept drift) was expected to be challenging, in particular for statistical approaches like binary relevance, because concepts had to be predicted without corresponding training data (cf. Table 1). Hence, a drop in performance was anticipated for these methods on this data. Indeed, BRLR and BRSVM were considerably deteriorated ($F_1 < 0.28$ on $\mathcal{D}_{\text{cat}}$). Thankfully, fusion allowed to absorb a certain amount of this decrease while it could not be prevented completely by the current systems.

Among the different fusion configurations, it seems that BRLR+MAUI and BRSVM+MAUI are on par with each other, and outperformed BRLR+D. The effect of post-processing by transformation rules appeared to be small, although positive effects are indicated in Figure 8. We assume that the approach has further potential. Maybe the restrictions on rules to high-level categories were too strict.

Introspection of frequent errors is in line with previous studies: in particular frequently appearing con-

Table 3: Comparison of approaches (averaged over 5 test sets). Architecture: L=lexical, A=associative, F=fusion. Bold type: highest values per setting and metric. Superscript $^{\downarrow}$: significantly smaller than maximum (bold) value (paired t-test, $p < .05$). [32] (Copyright © IEEE, see footnote 5)

	Method		sample-based avg.			concept-based avg.			
Data	Name	Arch.	F_1	prec.	rec.	F_1	prec.	rec.	$\overline{\lvert\mathcal{Y}_{\text{pred}}\rvert}$
$\mathcal{D}_{\text{shuffle}}$	DICT	L	0.277$^{\downarrow}$	0.329$^{\downarrow}$	0.273$^{\downarrow}$	0.222$^{\downarrow}$	0.451$^{\downarrow}$	0.265$^{\downarrow}$	4.92
$\mathcal{D}_{\text{shuffle}}$	DICT.T	L	0.286$^{\downarrow}$	0.334$^{\downarrow}$	0.285$^{\downarrow}$	0.223$^{\downarrow}$	0.450$^{\downarrow}$	0.267$^{\downarrow}$	5.07
$\mathcal{D}_{\text{shuffle}}$	MONQ	L	0.307$^{\downarrow}$	0.381$^{\downarrow}$	0.285$^{\downarrow}$	0.245$^{\downarrow}$	0.475$^{\downarrow}$	0.285$^{\downarrow}$	4.41
$\mathcal{D}_{\text{shuffle}}$	MAUI	L	0.332$^{\downarrow}$	0.486$^{\downarrow}$	0.280$^{\downarrow}$	0.256$^{\downarrow}$	0.459$^{\downarrow}$	0.291$^{\downarrow}$	3.28
$\mathcal{D}_{\text{shuffle}}$	BRLR	A	0.391$^{\downarrow}$	0.632	0.318$^{\downarrow}$	0.206$^{\downarrow}$	**0.558**	0.181$^{\downarrow}$	2.69
$\mathcal{D}_{\text{shuffle}}$	BRSVM	A	0.394$^{\downarrow}$	0.617$^{\downarrow}$	0.326$^{\downarrow}$	0.208$^{\downarrow}$	0.510$^{\downarrow}$	0.187$^{\downarrow}$	2.90
$\mathcal{D}_{\text{shuffle}}$	RHACK	A	0.413$^{\downarrow}$	**0.633**	0.342$^{\downarrow}$	0.211$^{\downarrow}$	0.553$^{\downarrow}$	0.190$^{\downarrow}$	2.98
$\mathcal{D}_{\text{shuffle}}$	BRLR+D	F	0.392$^{\downarrow}$	0.395$^{\downarrow}$	0.436$^{\downarrow}$	0.279	0.426$^{\downarrow}$	0.351$^{\downarrow}$	6.55
$\mathcal{D}_{\text{shuffle}}$	BRLR+MAUI	F	0.449$^{\downarrow}$	0.521$^{\downarrow}$	0.439$^{\downarrow}$	0.303	0.433$^{\downarrow}$	0.366$^{\downarrow}$	4.91
$\mathcal{D}_{\text{shuffle}}$	BRLR+MAUI.T	F	**0.449**	0.521$^{\downarrow}$	0.439$^{\downarrow}$	**0.303**	0.433$^{\downarrow}$	0.367$^{\downarrow}$	4.91
$\mathcal{D}_{\text{shuffle}}$	BRSVM+MAUI	F	0.449	0.512$^{\downarrow}$	0.444$^{\downarrow}$	0.300$^{\downarrow}$	0.417$^{\downarrow}$	0.369$^{\downarrow}$	5.08
$\mathcal{D}_{\text{shuffle}}$	BRSVM+MAUI.T	F	0.449	0.512$^{\downarrow}$	**0.445**	0.300$^{\downarrow}$	0.416$^{\downarrow}$	**0.370**	5.09
$\mathcal{D}_{\text{cat}}$	DICT	L	0.292	0.344	0.285$^{\downarrow}$	0.206$^{\downarrow}$	0.420	0.261$^{\downarrow}$	5.29
$\mathcal{D}_{\text{cat}}$	DICT.T	L	0.300$^{\downarrow}$	0.349$^{\downarrow}$	0.298$^{\downarrow}$	0.208	0.418	0.263$^{\downarrow}$	5.47
$\mathcal{D}_{\text{cat}}$	MONQ	L	0.320$^{\downarrow}$	0.393$^{\downarrow}$	0.295$^{\downarrow}$	0.225$^{\downarrow}$	0.441	0.279$^{\downarrow}$	4.76
$\mathcal{D}_{\text{cat}}$	MAUI	L	0.300$^{\downarrow}$	0.466$^{\downarrow}$	0.245$^{\downarrow}$	0.233$^{\downarrow}$	0.436	0.279$^{\downarrow}$	3.26
$\mathcal{D}_{\text{cat}}$	BRLR	A	0.273$^{\downarrow}$	0.524$^{\downarrow}$	0.202$^{\downarrow}$	0.150$^{\downarrow}$	**0.467**	0.139$^{\downarrow}$	2.21
$\mathcal{D}_{\text{cat}}$	BRSVM	A	0.277$^{\downarrow}$	0.510$^{\downarrow}$	0.210$^{\downarrow}$	0.151$^{\downarrow}$	0.425$^{\downarrow}$	0.146$^{\downarrow}$	2.42
$\mathcal{D}_{\text{cat}}$	RHACK	A	0.298$^{\downarrow}$	**0.536**	0.226$^{\downarrow}$	0.159$^{\downarrow}$	0.465	0.154$^{\downarrow}$	2.50
$\mathcal{D}_{\text{cat}}$	BRLR+D	F	0.350$^{\downarrow}$	0.365$^{\downarrow}$	**0.374**	0.235$^{\downarrow}$	0.377$^{\downarrow}$	0.316$^{\downarrow}$	6.57
$\mathcal{D}_{\text{cat}}$	BRLR+MAUI	F	0.366	0.469$^{\downarrow}$	0.332$^{\downarrow}$	0.253$^{\downarrow}$	0.388$^{\downarrow}$	0.326	4.56
$\mathcal{D}_{\text{cat}}$	BRLR+MAUI.T	F	0.371	0.472$^{\downarrow}$	0.339$^{\downarrow}$	**0.253**	0.388$^{\downarrow}$	0.326	4.60
$\mathcal{D}_{\text{cat}}$	BRSVM+MAUI	F	0.366	0.458$^{\downarrow}$	0.338$^{\downarrow}$	0.249$^{\downarrow}$	0.371$^{\downarrow}$	0.328	4.75
$\mathcal{D}_{\text{cat}}$	BRSVM+MAUI.T	F	**0.371**	0.461$^{\downarrow}$	0.344$^{\downarrow}$	0.249$^{\downarrow}$	0.371$^{\downarrow}$	**0.328**	4.79
$\mathcal{D}_{\text{series}}$	DICT	L	0.268$^{\downarrow}$	0.338$^{\downarrow}$	0.247	0.189	0.417$^{\downarrow}$	0.241$^{\downarrow}$	4.89
$\mathcal{D}_{\text{series}}$	DICT.T	L	0.277$^{\downarrow}$	0.343$^{\downarrow}$	0.259$^{\downarrow}$	0.191	0.417$^{\downarrow}$	0.245$^{\downarrow}$	5.05
$\mathcal{D}_{\text{series}}$	MONQ	L	0.293	0.390$^{\downarrow}$	0.255	0.206$^{\downarrow}$	0.445$^{\downarrow}$	0.256$^{\downarrow}$	4.32
$\mathcal{D}_{\text{series}}$	MAUI	L	0.308	0.500$^{\downarrow}$	0.244	0.222$^{\downarrow}$	0.464$^{\downarrow}$	0.264$^{\downarrow}$	3.11
$\mathcal{D}_{\text{series}}$	BRLR	A	0.387$^{\downarrow}$	0.663	0.304$^{\downarrow}$	0.218$^{\downarrow}$	**0.639**	0.205$^{\downarrow}$	2.70
$\mathcal{D}_{\text{series}}$	BRSVM	A	0.389$^{\downarrow}$	0.645$^{\downarrow}$	0.312$^{\downarrow}$	0.224$^{\downarrow}$	0.598$^{\downarrow}$	0.217$^{\downarrow}$	2.87
$\mathcal{D}_{\text{series}}$	RHACK	A	0.409$^{\downarrow}$	**0.665**	0.327$^{\downarrow}$	0.229$^{\downarrow}$	0.628$^{\downarrow}$	0.223$^{\downarrow}$	2.99
$\mathcal{D}_{\text{series}}$	BRLR+D	F	0.394$^{\downarrow}$	0.416$^{\downarrow}$	0.413	0.259$^{\downarrow}$	0.435$^{\downarrow}$	0.344$^{\downarrow}$	6.54
$\mathcal{D}_{\text{series}}$	BRLR+MAUI	F	0.448	0.556$^{\downarrow}$	0.413$^{\downarrow}$	0.284	0.467$^{\downarrow}$	0.354$^{\downarrow}$	4.79
$\mathcal{D}_{\text{series}}$	BRLR+MAUI.T	F	**0.449**	0.556$^{\downarrow}$	0.414$^{\downarrow}$	0.284	0.467$^{\downarrow}$	0.355$^{\downarrow}$	4.80
$\mathcal{D}_{\text{series}}$	BRSVM+MAUI	F	0.447	0.544$^{\downarrow}$	0.418$^{\downarrow}$	0.285	0.454$^{\downarrow}$	0.362$^{\downarrow}$	4.95
$\mathcal{D}_{\text{series}}$	BRSVM+MAUI.T	F	0.447	0.544$^{\downarrow}$	**0.419**	**0.285**	0.454$^{\downarrow}$	**0.363**	4.96

ceptual descriptors are detected better by statistical methods than by lexical approaches. Hence, application of statistical methods may seem to be preferable at first sight, when only looking at micro-averaged F1 values. In general, certain applications like information retrieval, however, especially require that also rare concepts are detected in order to improve subject access, since infrequent descriptors often represent more distinctive aspects of the content of the document, and therefore have stronger discriminative power. Our results indicate that detection of rare concepts can be accomplished by careful combinations of associative methods with lexical approaches, following the fusion rationale.

Our experiments gave an impression on how the approaches may behave in practical settings when methods are applied to new domains. They are in line with our expectations from the analysis of system architectures. Similar to results from Jimeno-Yepes et al. [13], we also found improvements by meta-learning according to specific concepts (RHACK). In our setting, it was, however, considerably affected by concept drift.

A direct comparison to figures reported on full-texts in the economics domain [11] (micro-avg. $F_1 = .39$, based on random order) is difficult because our results base on less training data (ours: $\approx$ 20k vs. theirs: $>$ 60k) and short text (titles and author-keywords). In general, multiple factors influence the absolute performance including data set characteristics and the calculation of

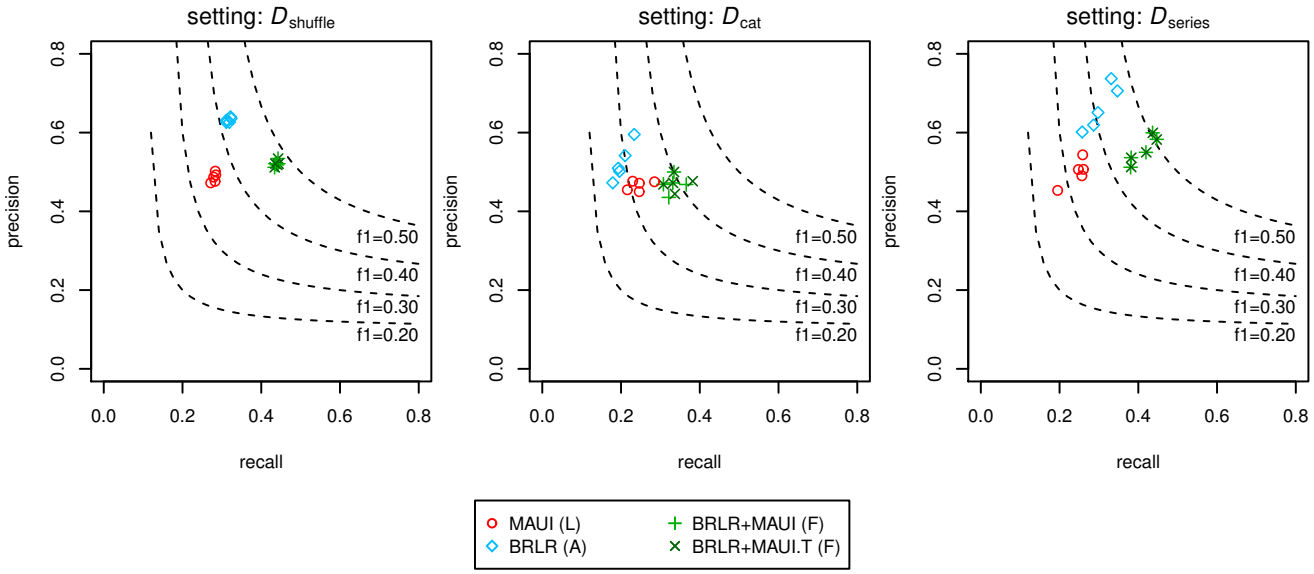


Fig. 7: Sample-based average precision and recall for random data set splits (left), explicit concept drift (center), implicit concept drift (right). Colors encode architectures (Lexical = red, Associative = blue, Fusion = green), symbols encode individual systems – Figure is best viewed in color.

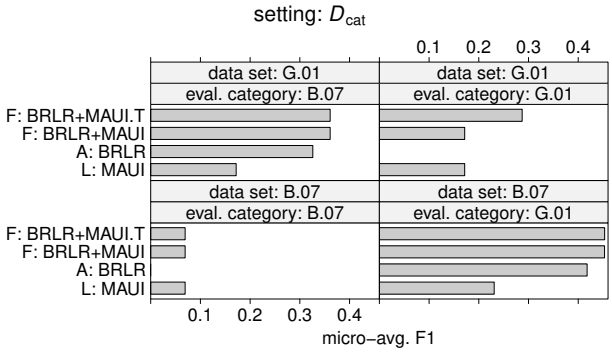


Fig. 8: Constrained evaluation on two folds (top: G.01, bottom: B.07) of $\mathcal{D}_{\text{cat}}$ (explicit concept drift) restricted to concepts of the corresponding categories (left, right) showing effects of system architecture and transformation rules. (Copyright © IEEE, see footnote 5)

F_1 scores (i.e., the type of averaging). Finally, please note that even professional indexers, which are commonly used as ground-truth [17, 13, 36, 11], do not agree on all indexing terms. For instance, Medelyan and Witten [20] reported an inter-indexer agreement of 39%. Albeit their values are not directly comparable to our work because of differences in data sets, thesauri, and indexing rules, they provide a rough overall impression. Determining upper bounds for indexing settings as performed by Medelyan and Witten is valuable but complex and costly, because it requires human indexing.

Besides comparing professionally indexed documents to automatically created results, collecting graded feedback with weights for individual concepts has been recognized as valuable for quality assessment of subject indexing [27]. The following section reports on recent efforts regarding such an evaluation of a fusion system.

8 Case Study with Graded Quality Assessment

Based on the experimental results and theoretical considerations reported in the previous sections, ZBW's automatic subject indexing group decided to pursue working with the fusion methodology. This section provides insights into ongoing efforts and results in this project.

At first it should be noted that weighting precision versus recall depends on application specific considerations. For the sake of generality, Table 3 listed results regarding precision, recall and their harmonic mean. In the special case of subject indexing, the relevancy of index terms can be measured in more detail using weights [27]. In the project regarding this case study, it has been specified that especially harmful descriptor assignments, that is, extremely irrelevant concepts, have to be avoided, which implicates that precision should be preferred to recall. As described in Section 4.2 and Section 7, it should be assumed that concept drift will

be present, which may disturb operations. In order to increase robustness, implementing a voting scheme in the fusion layer seemed reasonable, hence, the number and type of algorithms as well as the fusion function have been adjusted.

On the basis of our results, some of the implementations that performed worse than others having the same type of architecture have been discarded (for example, DICT and DICT.T). A simple kNN (k-nearest-neighbor with $k = 1$) based system has been added, which was expected to perform well when document titles are very similar. This method should, however, be handled carefully because harmful descriptors may also be retrieved, especially since it operates on titles and author keywords only. Consequently, additional filters have been applied. Based on the insights from our analysis reported in Section 5, special attention has been paid to the types of architectures available for the fusion layer and descriptor-invariance of the fusion function. In total, four individual systems were combined: two lexical (MAUI and STWFSA, which is an extension[22] of MONQ) and two statistical approaches (kNN and BRLR trained on titles). Their predictions have been merged by consecutive application of the following two fusion rules, R1 and R2, aiming at balancing precision and recall:

R1: Each concept proposed by fusion has to be supported by at least two individual methods. (voting scheme)

R2: Furthermore, documents are only considered when the number of concepts belonging to the union of categories economics (code: V) and business economics (code: B) is at least two.

The first rule addresses confidence in each assigned concept. The second rule aims to increase recall at the document level.

Mainly, the study was performed to assess the quality of the results produced by the fusion system which uses R1 and R2. Moreover, we were interested in the contributions of the four individual systems, which may enable improved fusion functions, hence all concepts suggested by at least one of them were examined. All systems were applied on the aforementioned new data (cf. Section 4.2), which has not been indexed by humans with STW concepts before. Detailed quality judgements have been collected using a web-based tool [33]. A group of 6 human indexers rated samples at the document level on a 3-point scale (reject, fair, good) and at the concept level on a 4-point scale (harmful, fair, helpful and really helpful). In total, 503 document reviews (on 454 distinct documents) have been entered.

Regarding the document level, the sets of index terms of the fusion system using R1 and R2 have been accepted (grades "good" and "fair") on more than three quarters of the documents. When documents were rejected, they did not contain more than one harmful descriptor in most cases.

[22] In STWFSA, we added special processing routines. For instance, it distinguishes upper and lower case words in certain cases, which in particular enables disambiguation of acronyms like SALT (Strategic Arms Limitation Talks) vs. salt (mineral) or AIDS (virus) vs. aids (plural of aid).

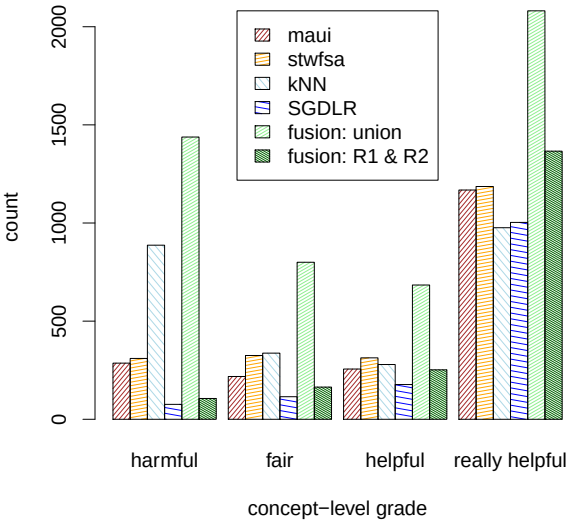


Fig. 9: Aggregated graded quality assessments on suggested concepts, showing how fusion with rules R1 and R2 exploits individual systems and balances precision and recall.

Table 4: Origin of concepts graded as really helpful.

Freq	support	kNN	MAUI	SGDLR	STWFSA
409	2		X		X
335	1	X			
254	4	X	X	X	X
249	3		X	X	X
216	2	X		X	
185	1			X	
107	1				X
97	3	X	X		X
93	1		X		
34	2		X	X	
28	2			X	X
25	3	X		X	X
20	2	X	X		
17	2	X			X
12	3	X	X	X	

Figure 9 depicts how often each grade[23] has been assigned to concepts for each method. Providing more detail, Table 4 itemizes how many really helpful concepts were found by distinct method combinations, where support denotes the number of systems that suggested these concepts. The majority of harmful and irrelevant concepts could be excluded by application of rules R1 and R2, while many helpful subject terms have been kept. Moreover, fusion proves to be highly relevant. As can be seen in the first row of Table 4, lexical systems (MAUI + STWFSA) triggered the assignment of 409 highly relevant concepts that none of the statistical associative systems detected, and they proposed more really helpful concepts than the statistical associative approaches (cf. Fig. 9). Comparison of both fusion functions (union vs. R1 & R2) on "really helpful" concept assignments in Figure 9 as well as detailed analysis of the potential (rows with support < 2) included in Table 4, it can be seen that there is still room for improvement by investigating techniques that leverage these already recognized relevant concepts and exclude harmful concepts at the same time.

In summary, the case study confirmed the value of fusion for automatic subject indexing. The additional filtering rules R1 and R2 yielded conservative assignments and allowed to favor precision over recall.

9 Limitations and Future Work

Open research questions arise in different fields, encompassing temporal effects, the extent of available metadata, quality control, as well as inference and learning algorithms.

This article put emphasis on general differences between training data and test data regarding topics and subject areas rather than considering language evolution explicitly. Further analysis and experiments on temporal phenomena like word sense changes (cf. [12, 30]) and their effects in the context of automatic subject indexing and scientific publications in economics remain open issues for future work.

In this article, we applied the approach to short texts (title + author keywords). It would be interesting to conduct more detailed experiments on the fusion methodology that compare different levels of extent of available text about the documents. In particular, we plan to investigate integration of abstracts, while still maintaining special consideration of titles, extending the approach of Jimeno-Yepes et al. [13].

[23] 49 documents have been rated by two indexers. Corresponding concept-level ratings have been averaged, using the floor function in order to resolve odd values.

Quality control is an ongoing issue which should be addressed continually in order to guarantee high quality subject terms for use in productive information retrieval systems. Measuring quality of automatic subject indexing appropriately is complex and can involve considerable costs (cf. Section 7, Section 8 and [27]). Notably, we are currently working on methods to estimate the quality of results automatically. Similar to evaluation based on document layout analysis for information extraction [34], we are planning to exploit different types of features and meta information such as membership in series and journals for automatic subject indexing.

As we have pointed out, descriptor-invariant learning is essential for subject indexing with respect to many aspects such as zero-shot learning, hence further research on this topic is worthwhile. Regarding lexical systems, further development may incorporate contextual markers and segmentation rules, which have been successfully applied in other domains, for instance, in terminology-driven clinical information extraction [34]. With respect to this, distributed word representations may be further explored to provide context for disambiguation, however, substantially new integration approaches may be necessary to reach more human-like concept learning [16], which we believe will be necessary for automatic subject indexing in the long term. Leveraging transformation-based learning as presented in Section 6 can be regarded as a step in this direction, adding a semantic learning and processing layer that generalizes across groups of concepts to the overall architecture. In agreement with discussions in the computational linguistics community [19], thinking about problems, architectures, and settings has to accompany exploration of different techniques and configurations.

10 Conclusion

In this article, we studied concept drift as a relevant and challenging issue in subject indexing for digital libraries. Based on an analysis of related work, we distinguish explicit and implicit concept drift, which in automatic subject indexing translates to settings with new descriptor terms and new types of documents, respectively. A theoretical analysis underlines that the system architecture is essential for the success of automatic subject indexing systems in settings with concept drift. Therefore, we proposed descriptor-invariant fusion of associative and lexical indexing approaches. Experiments in the domain of economics on texts, shorter than abstracts, showed that our fusion approach is superior to state-of-the-art methods for lexically-based and associative indexing. Fusion improved F_1 scores, in particular, when explicit or implicit concept drift was

induced by design of the training and testing data sets. In line with our initial considerations, superior F_1 values can be mainly attributed to substantial increases of recall. We also found positive effects of fusion in a case study, which supported the German National Library of Economics (ZBW) – Leibniz Information Centre for Economics to find suitable solutions for their practical setting.

Acknowledgements We thank all reviewers for their constructive advice. Moreover, we would also like to thank the indexing experts of the ZBW for valuable discussions and their support in gathering data for the experiments.

References

1. Aronson AR, Demner-Fushman D, Humphrey SM, Lin JJ, Ruch P, Ruiz ME, Smith LH, Tanabe LK, Wilbur WJ, Liu H (2005) Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. In: Voorhees EM, Buckland LP (eds) Proc. Text REtrieval Conference, TREC 2005, NIST, vol Special Publication 500-266
2. Bornmann L, Mutz R (2015) Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology 66(11):2215–2222
3. Breiman L (1996) Bagging predictors. Machine Learning 24(2):123–140, DOI 10.1007/BF00058655
4. Brill E (1995) Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics 21(4):543–565
5. Erbs N, Gurevych I, Rittberger M (2013) Bringing order to digital libraries: From keyphrase extraction to index term assignment. D-Lib Magazine 19(9/10), DOI 10.1045/september2013-erbs
6. Ferber R (1997) Automated indexing with thesaurus descriptors: A co-occurrence based approach to multilingual retrieval. In: Peters C, Thanos C (eds) Research and Advanced Technology for Digital Libraries, Springer, pp 233–252, DOI 10.1007/bfb0026731
7. Frank E, Paynter GW, Witten IH, Gutwin C, Nevill-Manning CG (1999) Domain-specific keyphrase extraction. In: Dean T (ed) Proc. Intl. Joint Conference on Artificial Intelligence, IJCAI '99, Morgan Kaufmann, pp 668–673
8. Gama J, Žliobaite I, Bifet A, Pechenizkiy M, Bouchachia A (2014) A survey on concept drift adaptation. ACM Computing Surveys (CSUR) 46(4):44
9. Gastmeyer M, Wannags M, Neubert J (2016) Relaunch des Standard-Thesaurus Wirtschaft – Dynamik in der Wissensrepräsentation. Inf Wiss & Praxis 67(4):217–240, DOI 10.1515/iwp-2016-0039
10. Gibaja E, Ventura S (2015) A tutorial on multilabel learning. ACM Comput Surv 47(3):52:1–52:38, DOI 10.1145/2716262
11. Große-Bölting G, Nishioka C, Scherp A (2015) A comparison of different strategies for automated semantic document annotation. In: Proc. Intl. Conference on Knowledge Capture, K-CAP 2015, ACM, pp 8:1–8:8, DOI 10.1145/2815833.2815838
12. Jatowt A, Duh K (2014) A framework for analyzing semantic change of words across time. In: IEEE/ACM Joint Conference on Digital Libraries, JCDL 2014, London, United Kingdom, September 8-12, 2014, IEEE Computer Society, pp 229–238, DOI 10.1109/JCDL.2014.6970173
13. Jimeno-Yepes A, Mork JG, Demner-Fushman D, Aronson AR (2012) A one-size-fits-all indexing method does not exist: Automatic selection based on meta-learning. JCSE 6(2):151–160, DOI 10.5626/JCSE.2012.6.2.151
14. Kessler J (2017) Scattertext: a browser-based tool for visualizing how corpora differ. In: Bansal M, Ji H (eds) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, System Demonstrations, Association for Computational Linguistics, pp 85–90, DOI 10.18653/v1/P17-4015
15. Kosnik LR (2015) What have economists been doing for the last 50 years? A text analysis of published academic research from 1960-2010. Economics: The Open-Access, Open-Assessment E-Journal 9:1–38, URL `http://dx.doi.org/10.5018/economics-ejournal.ja.2015-13`
16. Lake BM, Salakhutdinov R, Tenenbaum JB (2015) Human-level concept learning through probabilistic program induction. Science 350(6266):1332–1338
17. Lauser B, Hotho A (2003) Automatic multi-label subject indexing in a multilingual environment. In: Koch T, Sølvberg I (eds) Proc. Conf. Research and Advanced Technology for Digital Libraries, ECDL 2003, Springer, LNCS, vol 2769, pp 140–151, DOI 10.1007/978-3-540-45175-4_14
18. Loza Mencía E, Fürnkranz J (2010) Efficient multilabel classification algorithms for large-scale problems in the legal domain. In: Francesconi E, Mon-

temagni S, Peters W, Tiscornia D (eds) Semantic Processing of Legal Texts – Where the Language of Law Meets the Law of Language, LNAI, vol 6036, 1st edn, Springer, pp 192–215, DOI 10.1007/978-3-642-12837-0_11
19. Manning CD (2015) Computational linguistics and deep learning. Computational Linguistics 41(4):701–707, DOI 10.1162/COLI_a_00239
20. Medelyan O, Witten IH (2008) Domain-independent automatic keyphrase indexing with small training sets. Journal of the American Society for Information Science and Technology 59(7):1026–1040, DOI 10.1002/asi.20790
21. Medelyan O, Frank E, Witten IH (2009) Human-competitive tagging using automatic keyphrase extraction. In: Koehn P, Mihalcea R (eds) Proc. Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, ACM, pp 1318–1327
22. Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Pinker S, Nowak MA, Aiden EL (2010) Quantitative analysis of culture using millions of digitized books. Science 331(6014):176–182, DOI 10.1126/science.1199644
23. Nam J, Loza Mencía E, Kim HJ, Fürnkranz J (2015) Predicting unseen labels using label hierarchies in large-scale multi-label learning. In: Proc. Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2015, Springer, pp 102–118, DOI 10.1007/978-3-319-23528-8_7
24. Palatucci M, Pomerleau D, Hinton G, Mitchell TM (2009) Zero-shot learning with semantic output codes. In: Proc. Intl. Conference on Neural Information Processing Systems, NIPS '09, Curran Associates Inc., USA, pp 1410–1418
25. Pouliquen B, Steinberger R, Ignat C (2003) Automatic annotation of multilingual text collections with a conceptual thesaurus. Proc Workshop Ontologies and Information Extraction, EUROLAN 2003 abs/cs/0609059
26. Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND (eds) (2009) Dataset shift in machine learning. Neural information processing series, MIT Press, Cambridge, Mass, URL https://mitpress.mit.edu/books/dataset-shift-machine-learning
27. Rolling LN (1981) Indexing consistency, quality and efficiency. Information Processing & Management 17(2):69–76, DOI 10.1016/0306-4573(81)90028-5
28. Sappadla PV, Nam J, Loza Mencía E, Fürnkranz J (2016) Using semantic similarity for multi-label zero-shot classification of text documents. In: Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, d-side publications
29. Sebastiani F (2002) Machine learning in automated text categorization. ACM Computing Surveys 34(1):1–47
30. Tahmasebi N, Risse T (2017) On the uses of word sense change for research in the digital humanities. In: Kamps J, Tsakonas G, Manolopoulos Y, Iliadis LS, Karydis I (eds) Research and Advanced Technology for Digital Libraries - 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings, Springer, Lecture Notes in Computer Science, vol 10450, pp 246–257, DOI 10.1007/978-3-319-67008-9_20
31. Ting KM, Witten IH (1999) Issues in stacked generalization. J Artif Intell Res (JAIR) 10:271–289, DOI 10.1613/jair.594
32. Toepfer M, Seifert C (2017) Descriptor-invariant fusion architectures for automatic subject indexing. In: ACM/IEEE Joint Conference on Digital Libraries, JCDL 2017, Toronto, ON, Canada, June 19-23, 2017, IEEE Computer Society, pp 31–40, DOI 10.1109/JCDL.2017.7991557, URL https://doi.org/10.1109/JCDL.2017.7991557
33. Toepfer M, Seifert C (2017) Towards Semantic Quality Control of Automatic Subject Indexing, Springer International Publishing, Cham, pp 616–619. DOI 10.1007/978-3-319-67008-9_56, URL https://doi.org/10.1007/978-3-319-67008-9_56
34. Toepfer M, Corovic H, Fette G, Klügl P, Störk S, Puppe F (2015) Fine-grained information extraction from german transthoracic echocardiography reports. BMC medical informatics and decision making 15:91, DOI 10.1186/s12911-015-0215-x
35. Tsoumakas G, Laliotis M, Markantonatos N, Vlahavas IP (2013) Large-scale semantic indexing of biomedical publications. In: Ngomo AN, Paliouras G (eds) Proceedings of the first Workshop on BioMedical Semantic Indexing and Question Answering, CEUR-WS.org, CEUR Workshop Proceedings, vol 1094, URL http://ceur-ws.org/Vol-1094/bioasq2013_submission_6.pdf
36. Wilbur WJ, Kim W (2014) Stochastic gradient descent and the prediction of mesh for pubmed records. Proc AMIA Annual Symposium pp 1198–1207
37. Wolpert DH (1992) Stacked generalization. Neural Networks 5(2):241–259, DOI 10.1016/S0893-6080(05)80023-1