Inf Retrieval (2014) 17:520–544
DOI 10.1007/s10791-014-9239-6

INFORMATION RETRIEVAL IN THE INTELLECTUAL PROPERTY DOMAIN

Dealing with temporal variation in patent categorization

Eva D'hondt · Suzan Verberne · Nelleke Oostdijk · Jean Beney · Cornelius Koster · Lou Boves

Received: 22 March 2013 / Accepted: 28 February 2014 / Published online: 25 March 2014

Abstract In this paper, we quantify the existence of *concept drift* in patent data, and examine its impact on classification accuracy. When developing algorithms for classifying incoming patent applications with respect to their category in the International Patent Classification (IPC) hierarchy, a temporal mismatch between training data and incoming documents may deteriorate classification results. We measure the effect of this temporal mismatch and aim to tackle it by optimal selection of training data. To illustrate the various aspects of concept drift on IPC class level, we first perform quantitative analyses on a subset of English abstracts extracted from patent documents in the CLEF-IP 2011 patent corpus. In a series of classification experiments, we then show the impact of temporal variation on the classification accuracy of incoming applications. We further examine what training data selection method, combined with our classification approach yields the best classifier; and how combining different text representations may improve patent classification. We found that using the most recent data is a better strategy than static sampling but that extending a set of recent training data with older documents does not harm

Cornelius Koster—deceased.

E. D'hondt (✉) · S. Verberne · N. Oostdijk · C. Koster · L. Boves
Radboud University Nijmegen, Nijmegen, The Netherlands
e-mail: e.dhondt@let.ru.nl

S. Verberne
e-mail: s.verberne@cs.ru.nl

N. Oostdijk
e-mail: n.oostdijk@let.ru.nl

C. Koster
e-mail: kees@cs.ru.nl

L. Boves
e-mail: l.boves@let.ru.nl

J. Beney
Université de Lyon, Lyon, France
e-mail: jean.beney@ouvaton.org

classification performance. In addition, we confirm previous findings that using 2-skip-2-grams on top of the bag of unigrams structurally improves patent classification. Our work is an important contribution to the research into concept drift for text classification, and to the practice of classifying incoming patent applications.

Keywords Concept drift · Patent classification · Text representation

1 Introduction

Like most large-scale text corpora that are collected over a longer period of time, patent corpora are subject to temporal variation. The constant introduction of new technologies and their corresponding vocabularies in the various technical fields leads to shifts in the underlying distribution of words over time (Tsymbal 2004).

Let us illustrate with an example: Consider the category 'Telephonic Communication' (H04M).[1] In the 1970s a typical granted patent in this category could describe an automatic answering machine, complete with cassette deck, magnetic tape and turn buttons. A typical 2012 patent in the same category may cover a new type of smart phone with camera and a touch screen. Although new patents are still filed for answering machines, such systems will be digital and unlikely to contain many of the components of their 1970s predecessors. This example shows that the contents of the H04M category have evolved over time: It covers different concepts expressed with different words. This category-internal shift has corpus-wide consequences: In 2012 the 'Telephonic Communication' (H04M) category is more similar to the 'Electric Digital Data Processing' (G06F) category than it was in the 1970s. Another aspect of temporal change is the rise and decline of certain categories: The recent explosion of innovation in the smart phone industry has infused the H04M category with many patent application filings in the last decade, while other fields such as 'Methods for Organic Chemistry' (C07B) have experienced a slower innovation rate in the last few years, and consequently seen a decline in the number of patent applications per year.

Such shifts in the words that characterize a category and the relative size of categories in (text) collections are known as *concept drift*. The term was first introduced by Schlimmer and Granger (1986) and refers to a non-stationary learning problem over time (Žliobaitė 2009).

One way to model the text classification process is by means of the noisy channel model (Schlimmer and Granger 1986). This is a probabilistic model which assumes that the observed phenomena are generated by one or more hidden sources, and sent through a channel that may distort the signals. In our case, the observed phenomena are the documents and the hidden sources are characterized by the distribution of the features (words) that make up the documents. A source generates texts pertaining to one specific category or 'concept'. If the sources are stationary, the distribution of the features that characterize a

[1] H04M is a language-independent symbol in the International Patent Classification (IPC), a complex hierarchical classification system comprising sections, classes, subclasses, groups and subgroups. Each level in the hierarchy has a different granularity. (Certain aspects of) a smart phone would fall under section H *'Electricity'*, class 04 *'Electric Communication Technique'*, subclass M *'Telephonic Communication'*, group 1 *'Substation equipment'*, subgroup 725 *'Cordless Telephone'*. The revised version of the IPC (IPC8) used in the CLEF-IP 2011 corpus contains eight sections, 129 classes, 639 subclasses, 7,352 groups and 61,847 subgroups (Benzineb and Guyot 2011). The IPC covers all technological fields in which inventions can be patented. In the experiments reported in this article we opted to classify on class level.

concept is stable over time, so that sets of training and test documents would have the same feature distribution, irrespective of the time when the documents were generated. In real-life classification tasks, however, we see (systematic[2]) changes in the sources over time, so that a classifier trained with texts generated by the source at time T_1 may not match very well with text generated at time T_2. All research focussed on concept drift basically has the same goal: Reducing the impact of the mismatch between the feature distributions in the train and in the test data.

As first proposed by Kelly et al. (1999), the presence of concept drift may be observed in three ways:

1. The *distributions of category labels*, i.e. the (relative) number of documents assigned to the different categories, generated by the various sources, may change over time. In the example given previously this refers to the growth of the H04M category over time relative to other categories.
2. The *term distributions that characterize the sources* may change over time when new terms are introduced and older terms become obsolete, for example the introduction of the phrase ‘touch screen’ in the H04M category.
3. Through shifts in term distributions the *source similarity* between categories may also change over time, e.g. the H04M and G06F categories have become more similar over time.

Concept drift has received a lot of interest in the last decades in various research fields dealing with large amounts of (incoming) data (Žliobaitė 2009), such as recommender systems (Koychev 2000), adaptive information filtering of news feeds (Lebanon and Zhao 2008), classification of scientific articles (Mourão et al. 2008), etc., but to our knowledge concept drift has received almost no attention in the context of (English) patent classification. In fact, most of the current research on improving automated patent classification generally treats corpora of patent documents as static wholes, where the training data distribution and the test data distribution are similar.

This approach is naive at best: The patent domain is prone to change, since a patent (application) can only be granted if it brings a novel element or implementation to its technical field(s). In a real-life setting there is often a mismatch between the test data (incoming patent applications), and the training data available in the corpus, especially in fast-paced categories where rapid innovation occurs. In previous work (Verberne et al. 2010), we noticed that a two-year gap between the training data and the test document set can already cause a large drop in classification accuracy. In this article we will investigate this phenomenon in more detail, while taking some of the unique properties of the patent domain and patent classification task into account.

First, patent classification is a multi-label, highly imbalanced classification problem: In the CLEF-IP 2011 patent corpus, which was used for this article, 20 % of the categories comprise 80 % of the documents. For a large subset of the categories few new patent applications are submitted, which may mean less temporal variation, but surely means little training data. Where most patent classification research has focussed on improving the classification accuracy of the larger classes, this is not defensible in a real-life setting: Incoming patent applications must be routed to the correct examiner(s), no matter how small

[2] A difficult problem in handling concept drift is distinguishing between true concept drift and random noise. In the beginning of gradual drift (see Sect. 2), when only few instances generated by the new version have been seen, it is difficult to distinguish between random noise and a genuine change in the data distribution that characterizes the source.

the relevant categories may be. It is therefore important to investigate whether concept drift has a different impact in large and small categories. Please note that in this paper we only study temporal variation at class level. While there is a demand to classify on lower levels in the IPC hierarchy (Benzineb and Guyot 2011), the data sparseness in the subgroup categories poses problems for low-level classification even with static sampling. Spreading out the training data with respect to different time stamps would create sparseness problems which would render it impossible to draw conclusions on temporal variation.

Second, the language use in patents is quite different from most other text genres. Patents are written in so-called *patentese*: a version of English with long sentences in complex syntactic constructions, full of genre-specific formulations and using a large vocabulary of Multi-Word Terms.[3] In previous research (D'hondt et al. 2012, 2013) we found that for classifying abstracts[4] of patent applications written in English, extending a bag-of-words representation with phrasal features significantly improves classification accuracy, as phrases can capture the most important Multi-Word Terms. However, in these experiments we completely ignored concept drift. We suspect that phrasal features—perhaps even more than words—can be subject to temporal variation and therefore careful selection of the phrasal features may improve classification of incoming patent applications.

Using a subset of data from the CLEF-IP 2011 patent corpus, consisting of 360,000 English patent abstracts dating from 1981 to 2004 that had at least one IPC category label on class level, we will investigate:

1. How concept drift is manifested in the corpus;
2. Whether classification accuracy improves when using temporally-aware sampling, compared to static sampling[5] in training the classifiers;
3. What the optimal trade-off is between recency and training window size when selecting training data;
4. Whether the finding in other work that adding phrasal features significantly improves classification still holds for our temporally-sensitive selection of training data.

In this article we will present quantitative analyses of the distribution of the categories and their characteristic terms (features) as well as results from classification experiments. From these results we will draw conclusions on what data selection methods are most suitable for training patent classifiers. We expect that these insights will be of interest to the patent community as a whole, as well as to other prospective researchers who want to examine automated text classification on large, imbalanced data sets that were collected over a long time period.

The remainder of this article is structured as follows. Section 2 describes related work on concept drift in text classification tasks. In Sect. 3 we present general information on the corpus and the classification algorithm used in our experiments and analyses. Section 4 illustrates the three ways in which concept drift occurs in the CLEF-IP 2011 patent corpus

[3] A Multi-Word Term (MWT) is a term that is composed of more than one word. The exact semantics of a Multi-Word Term differ per knowledge area and cannot be inferred directly from its parts (SanJuan et al. 2005; Frantzi et al. 1998).

[4] (Full) patent documents consist of four different sections, e.g. the title, abstract, claims and description section, each with their own particular language use. We opted to only use abstracts as they are the easiest to process and contain the most concise descriptions of the inventions patented. We imagine our findings may easily extend to the other sections of the document.

[5] The term 'static sampling' used in this article reflects this stationarity assumption. It refers to a method of dividing a document corpus in a train and test set without taking the time stamps of the documents into account.

(subset). In Sect. 5 we investigate the impact of concept drift on patent classification accuracy. Section 6 examines the trade-off between the recency effect and training window size. In Sect. 7 we examine the effect of adding phrasal features. Concluding remarks are given in Sect. 8.

2 Related work: concept drift in text classification

In this overview, we will limit ourselves to research on concept drift done in the context of text classification. For a detailed overview on concept drift spanning multiple fields of research, please see the excellent introduction by Žliobaitė (2009) and a shorter overview by Tsymbal (2004).

In the literature, most researchers distinguish between three types of drift, depending on the rate and periodicity of the change.

1. *Sudden concept drift*:[6] In sudden concept drift there is a clear moment when the distribution in the corpus changes substantially. The primary aim of systems dealing with sudden drift is to accurately detect shifts and react according to this trigger by changing the training set selection to only include the instances that match the new distribution, and retraining the systems. Typical examples where sudden concept drift plays a role are adaptive filtering (Scholz and Klinkenberg 2007) where users' interest may change suddenly and spam filtering (Fawcett 2003), where spammers actively try to get around existing spam filtering systems.
2. *Gradual concept drift*: In gradual concept drift the data distribution in the corpus changes continuously over time, though not necessarily at a constant rate. Consequently, there is no clear event that signals change. It has been suggested that gradual drift is best handled by moving windows (of fixed size) on the data (Kuncheva 2004). Prime examples of real-life gradual drift in textual data can be found in email categorization (Carmona-Cejudo et al. 2011) and classification of scientific articles in the ACM Digital Library and the Medline medical collections (Mourão et al. 2008).
3. *Recurring drift*: In recurring concept drift, an older version of a source, which recently was less active, can suddenly become more prominent, thus changing the data distribution back to an earlier state. For recurring concept drift problems, a suitable technique is to keep old classifiers in store and measure their performance on incoming test data. Recurring drift can be found in news classification (Forman 2006).

In text classification applications concept drift is often handled by applying one or more of the following techniques:

- *instance selection*, which entails selecting parts of the training data that are relevant to the current status of the source. This is most commonly done by sliding time windows of fixed or adaptive size over the most recently arrived instances. Windows of adaptive size are usually determined by so-called 'triggers', i.e. changes in the training set distribution over time, or drops in classification accuracy on incoming test data;
- *instance weighting*, which refers to approaches that use the ability of some classification algorithms like Support Vectors Machines to give some training documents more weight during training (Klinkenberg 2004);
- combining classifiers trained on different (training) data sets in *dynamic ensembles*.

[6] Also known as 'concept shift'.

The choice of technique(s) depends on the type of drift in the corpus. In the following paragraphs we will give an overview of some of the work done on concept drift in various text classification tasks and show how the drift type and task characteristics determine which techniques are best used. Please note that we will only focus on those tasks where temporal variation takes place on content level, that is, in changes in the distribution of words, overlap between categories, etc., rather than in the form of changes in the interest of users: Much work has been done on tracking user interest in the context of improving adaptive information filtering, recommender systems, spam filtering etc. Although these problems also deal with corpora that contain temporal variation, the real shift lies in the (abrupt) changes in user interest, not in the intrinsic content changes in the corpora. Moreover, modelling user interest is often a one-class classification problem: Whether or not the content is *relevant* for a particular user. In contrast, we discuss classification research that deals with the intrinsic content changes within and between (multiple) categories.

Email classification is a multi-class classification problem with typically 10–100 categories, depending on the data set used. Emails are mostly fairly short text fragments and contain a lot of metadata such as sender information, time stamp, etc. Although the overall structure of an email folder hierarchy may change considerably over time, content-wise an email corpus tends to change more gradually: for example, in emails concerning a work-related project, collaborators may come and go but the project topic will not shift completely from one day to another. Segal and Kephart (1999) proposed an incremental learner that recalibrates the TF-IDF vectors of the various categories with each incoming email. As time goes by, some term features on the TF-IDF vectors may become obsolete, which results in increasingly lower TF-IDF values, but are never completely discarded. Rather than using the full training set, Carmona-Cejudo et al. (2011) implemented adaptive window and controlled forgetting techniques using the Drift Detection Method (DDM) (Ja et al. 2004), which signals drift based on the performance of the most recent model on incoming data. They found that using instance selection instead of all available data results in significant improvements on the ENRON email dataset (Klimt and Yang 2004).

News classification has significantly different properties than email classification. While the number of categories usually remains constant over time, the category content changes very fast. For something to be 'news' it must be different from what was relevant in the same category the day before. What sets news classification apart from other text classification tasks—like patent classification—is the fact that it has recurring themes. Consider the papal resignation in 2013: At the first mention on the 28th of February, it dominated international news for a couple of days. At that time point, term features like 'pope', 'vatican', 'religion', etc., were strong predictors of the 'international news' category in a news corpus. Media attention died down after a while and was only rekindled a week later when the next papal conclave was due to start. At that point the term features related with the papal resignation became relevant again for the 'international news' category. Lebanon and Zhao (2008) used the Reuters RCV1 dataset (Lewis et al. 2004) which consists of 800,000 documents spanning one year of new stories to track changes in class-internal feature distributions of the three most popular categories. Using models of the local likelihood of word appearance, they illustrated the existence of concept drift in the corpus, but do not specify which type. Forman (2006) used the same corpus to perform the Daily Classification Task, in which incoming news stories are classified into four categories, i.e. sports (GSPO), government and social issues (GCAT), economics (ECAT) and money markets (M13). It is assumed that only a fraction of the incoming text documents per day are labelled, while the rest of the documents are unlabeled and require automatic labeling.

Forman wants to leverage the knowledge inherent in older classification models, while at the same time giving the most weight to the most recent training models (i.e. that day's training set). This is achieved by expanding the feature vectors of the training documents with labels given to those documents by older classifiers. Although there is a clear improvement of these extended features when using oracle data, i.e. manually checked labels assigned by older classifiers, the real-life results show that these features cannot adequately deal with the data sparseness of the very small training sets. Šilić et al. (2012) performed experiments using a logistic regression classifier on a 248 K corpus, comprising seven categories, from the French newspaper *Le Monde* to illustrate concept drift. Using a much higher granularity (years rather than days) than Forman (2006) they discovered incremental (gradual) concept drift in the corpus.

A third multi-class text classification task is the automatic *classification of web documents*. This task differs from news classification in that the temporal dimension is less explicit in document creation: While news reports often build on facts from previous news reports, web documents more often are stand-alone descriptions of a certain topic. Liu and Lu (2002) experimented on a dataset of 1838 web documents extracted from yahoo.com, with 83 (hierarchical) categories in total on topics from 'science', 'computers and internet' and 'society and culture'. They present an adaptive classifier based on updating weights of terms that are most representative for a category. They find that an evolutionary maintenance of the feature sets is essential for keeping accuracy scores constant over time. Like Segal and Kephart (1999) they do not remove older terms from the models, but rather demote them.

The text classification task most closely related to patent classification is the *classification of technical documents*, which also contain large numbers of technical terms. A fair amount of work has been done on this topic by Mourão et al. (2008), Rocha et al. (2012), Salles et al. (2010) who focussed on concept drift in text classification on large collections of scientific and medical articles, i.e. the ACM Digital Library which contains 30 K articles over 23 years, divided in 11 categories; and the Medline dataset which contains 861 K articles over 16 years in seven categories, respectively. In Mourão et al. (2008) temporal aspects of the document collections are examined and quantified. The authors illustrate the existence of (gradual) concept drift and find that the optimal trade-off between recency and training set size is category-dependent. In Rocha et al. (2012) the authors describe the Chronos algorithm which performs example selection of document batches (per year) in the training data, based on the descriptiveness of features in training and test documents. In Salles et al. (2010), temporal weighting of examples and classification scores is incorporated in the Rocchio, k-NN, and Naive Bayes learning algorithms. Cohen et al. (2004) examines automated classification of biomedical documents in the TREC 2004 Genomics triage task. They notice a clear drop in classification accuracy between the cross-validation results on the training set and the test accuracy, which is caused by concept drift. Further analysis shows that the concept drift is not caused by an influx of new terms in the more recent data, but by shifts in category overlap.

To our knowledge, only one previous study has explicitly investigated temporal variation in patent classification (Ma et al. 2009) although the effect of temporal gaps between training and test material has also been noticed in CLEF-IP contributions (Verberne et al. 2010). In the context of the classification and retrieval tracks organised by the Japanese patent office (Nanba et al. 2008, 2010), Ma et al. (2009) evaluated temporal differences in Japanese (full) patents from the NTCIR-5 patent data set[7] which contains 2.4 million patents from 1993 to

[7] The data set can be obtained at http://research.nii.ac.jp/ntcir/permission/ntcir-5/perm-en-PATENT.html.

1999. They find that vocabulary use in two patents from one category is much more similar if the patents are close to each other in time than when there is a large time gap between them, and that newly introduced terms in recent years are most likely domain-specific terms. The authors then propose an approach using min-max-modular Support Vector Machines, in which (prior) knowledge on meta-data like the time stamps and IPC class labels of patent documents is used to decompose the classification task into a series of two-class subproblems. The temporally-aware version of their algorithm effectively splits up training data per year. The algorithm then creates an ensemble of two-class subclassifiers on these batches and uses these to score the most recent documents in the 1998–1999 test set. The scores from the various subclassifiers are then mapped onto one score per category for each document. They find that selecting training documents on the basis of time stamp outperforms static, i.e. temporally unaware, selection. However, the biggest improvement in classification accuracy is achieved when the classification problem is split into subproblems based on IPC category information[8] in addition to temporal information. It should be noted they define patent classification as a mono-label classification task, and classify at the highest level of the IPC hierarchy, i.e. *sections* (eight categories) only.

3 Method

3.1 Winnow algorithm

The classification experiments presented in this paper were carried out within the framework of the Linguistic Classification System (LCS) (Koster et al. 2003).[9] The LCS has been developed for the purpose of comparing different text representations. Currently, three classifier algorithms are available in the LCS: Naive Bayes, Balanced Winnow (Dagan et al. 1997), and SVM-light (Joachims 1999). Koster and Beney (2009) found that Balanced Winnow and SVM-light give comparable classification accuracy scores for patent texts on a data set similar to ours, but that Winnow is much faster than SVM-light for classification problems with a large number of features and categories. We therefore only used the Balanced Winnow algorithm for our classification experiments, which were run with the following LCS configuration, based on tuning experiments on data from the same corpus, by Koster et al. (2011) and D'hondt et al. (2013):

- Global term selection (GTS): Document frequency minimum is 2, term frequency minimum is 3. Although initial term selection is necessary when dealing with such a large corpus, we deliberately aimed at keeping as many of the sparse phrasal terms as possible.
- Local term selection (LTS): We used the simplified Chi Square, as proposed by Galavotti et al. (2000) to automatically select the most representative terms for every category, with a hard maximum of 10,000 terms per category.[10]

[8] This corresponds to a hierarchical classification approach.

[9] http://www.phasar.cs.ru.nl/LCS/.

[10] In D'hondt et al. (2013), we found that increasing the cut-off to 100,000 terms resulted in a small increase in accuracy (F1 values) for the combined representations, mostly for the larger categories. Because the patent domain has a large lexical variety, a large amount of low-frequency terms in the tail of the term distribution can have a large impact on the accuracy scores. Since we are more interested in the relative gains between different text representations and the corresponding top terms in the category profiles, than in achieving maximum classification scores, we opted to use only 10,000 terms for efficiency reasons.

- After LTS, the selected terms of all categories are *aggregated* into one combined term vocabulary, comparable to the Round-Robin strategy proposed by Forman (2004). This combined vocabulary is used as the starting point for training the individual categories.
- Term strength calculation: LTC algorithm (Salton and Buckley 1988) which is an instance of the TF-IDF measure.
- Training method: Ensemble learning based on one-versus-rest binary classifiers.
- Winnow configuration: We used the same setting as Koster et al. (2011), D'hondt et al. (2013), namely $\alpha = 1.02$, $\beta = 0.98$, $\theta+ = 2.0$, $\theta- = 0.5$ with a maximum of 10 training iterations.
- For each document the LCS returns a ranked list of all possible category labels and the corresponding Winnow scores. If the score assigned to a category is higher than a predetermined threshold, the document is assigned that category. We used the natural threshold of the Winnow algorithm equal to one. We configured the LCS to return a minimum of one label (with the highest score, even if it is lower than the threshold) and a maximum of four labels for each document. These values are based on the average number of classes per patent in the training data.

3.2 Evaluation measures

The classification quality of the various classifiers was evaluated using the F-measure metric (F1), which is equal to the harmonic mean of recall and precision. The F1 score can be computed in two ways: *Micro*-averaged or *Macro*-averaged. The micro-averaged score is an average over all the document-category tuples, in which each document is given equal weight. Given the data imbalance in the patent corpus, micro-averaged F1 scores will give us insight in the performance of the larger categories. The macro-averaged F1 score, on the other hand, gives equal weight to each category and is therefore a good measure to see classifier performance on the smaller categories.

3.3 Corpus selection

We chose to use the CLEF-IP 2011 corpus as it is a large-scale patent corpus which spans several decades. From the corpus we extracted all English abstracts from documents that have one or more category labels on IPC class level. This led to a total of 1,004,022 English abstracts. For each document we also extracted the unique filing date (priority date) of the patent application in the system. We opted for a fine granularity, i.e. binning the documents per year, because this enabled us to examine the differences in drift rates between the various categories in more detail.

The distribution of the number of documents as a function of time in the corpus is shown in Fig. 1. There is a clear imbalance in the number of documents available for different years. In order to avoid a bias for a specific time period, we opted to select an equal number of documents for each year. This resulted in a subset from the data ranging from 1981 up to 2004, depicted by the (red) box in Fig. 1. The subset is divided into different batches, each batch containing 15,000 documents and spanning one year. Each batch is sampled according to the category label distribution for that year in the corpus. In total, the subcorpus used in the remainder of this paper consists of 360,000 documents.

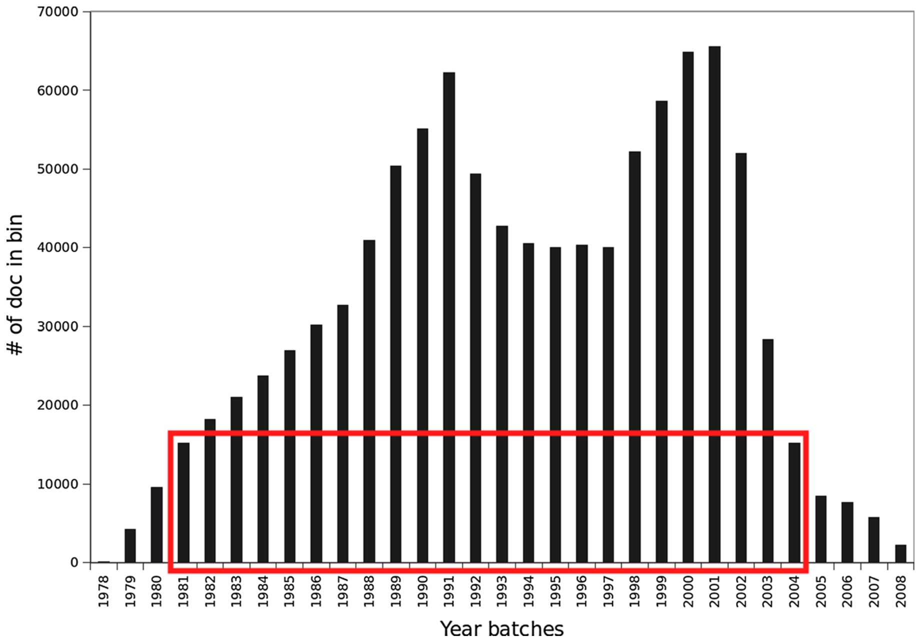

Fig. 1 Data distribution of English abstracts in CLEF-IP 2011 corpus, in # of documents per year. The (*red*) *box* shows the data subset used in subsequent analyses and classification experiments (Color figure online)

3.4 Preprocessing

General preprocessing of the extracted texts in the subcorpus included removing XML tags, cleaning up character conversion errors and removing references to claims, image references and in-text list designators from the original texts. This was done automatically using regular expressions. We then ran a Perl script to divide the running text into sentences, by splitting on end-of-sentence punctuation such as question marks and full stops. In order to minimize incorrect splitting of the texts, the Perl script was supplied with a list of common English abbreviations and a list containing abbreviations and acronyms that occur frequently in technical texts,[11] derived from the Specialist lexicon.[12]

Previous research (D'hondt et al. 2013) has shown that patent classification accuracy increases when unigram features are combined with phrasal features in a bag-of-words approach. In D'hondt et al. (2012) we found that filtering the resulting features on Part-of-Speech (PoS) information, i.e. only allowing nouns, verbs or adjectives (or combinations thereof), significantly raises the performance compared to using all phrasal features. In the case of unigrams, PoS filtering has no significant effects on classification performance, but it does result in a smaller (and more manageable) feature set. We found that combining both representations yielded the best classification accuracy scores.

The preprocessed sentences were therefore tagged using an in-house PoS tagger (van Halteren 2000).[13] The tagger was trained on the annotated subset of the British National

[11] Both the splitter and abbreviation file can be downloaded from https://sites.google.com/site/ekldhondt/downloads.

[12] The lexicon can be downloaded at http://lexsrv3.nlm.nih.gov/Specialist/Summary/lexicon.html.

[13] Tokenization was performed by the tagger.

Table 1 Example output: PoS-filtered words and PoS-filtered 2-skip-2-grams for the expression *'processes for preparing said compound'*

Words (PoS-filtered)		2-Skip-2-grams (PoS-filtered)	
process	(N)	process_prepare	(N_V)
prepare	(V)	process_say	(N_A)
say	(A)	prepare_say	(V_A)
compound	(N)	prepare_compound	(V_N)
		say_compound	(A_N)

Corpus and uses the CLAWS-6 tag set.[14] We chose this particular tagger because it is highly customizable to new lexicons and word frequencies: Language usage in the patent domain can differ greatly from that in other genres. For example, the past participle *said* is often used to modify nouns as in *'for said claim'*. While this usage is very rare and archaic in general English where *said* most often occurs as a simple past tense or past participle, it is a very typical modifier in patent language. Consequently, for tagging text from the patent genre, a PoS tagger must be updated to account for these differences in language use, so as to output more accurate and better informed tags and tag sequences. For the following experiments, we have adapted the tagger to use word frequency information and associated PoS tags from the patent domain, taken from the AEGIR lexicon.[15] However, we have not retrained the tagger on any annotated patent texts. Such annotations are very expensive to make and were not possible within the scope of this article. Consequently, the tagger is still only trained on the label sequences in the original training texts, i.e. the British National Corpus.

From this data we generated two text representations using the filtering and lemmatisation procedure described in D'hondt et al. (2012): PoS-filtered words (only allowing nouns, verbs and adjectives), and PoS-filtered 2-skip-2-grams (only allowing combinations of nouns, verbs and/or adjectives). Through this process, for the phrase *'processes for preparing said compound'* the terms presented in Table 1 are generated.

4 Illustrating concept drift in the patent corpus

In this section we investigate the existence of concept drift in the CLEF-IP 2011 patent corpus by looking at the three ways in which drift may manifest itself: (a) changes in category distributions over time (Sect. 4.1), (b) category-internal feature shifts (Sect. 4.2), and (c) category similarity over time (Sect. 4.3). All analyses in the remainder of this paper are on the class level in the IPC taxonomy; therefore, we need to distinguish between 121 categories.

[14] http://ucrel.lancs.ac.uk/claws6tags.html.

[15] The AEGIR lexicon is part of the AEGIR parser, a hybrid dependency parser that is designed to parse technical text. For more information, see Oostdijk et al. (2010).

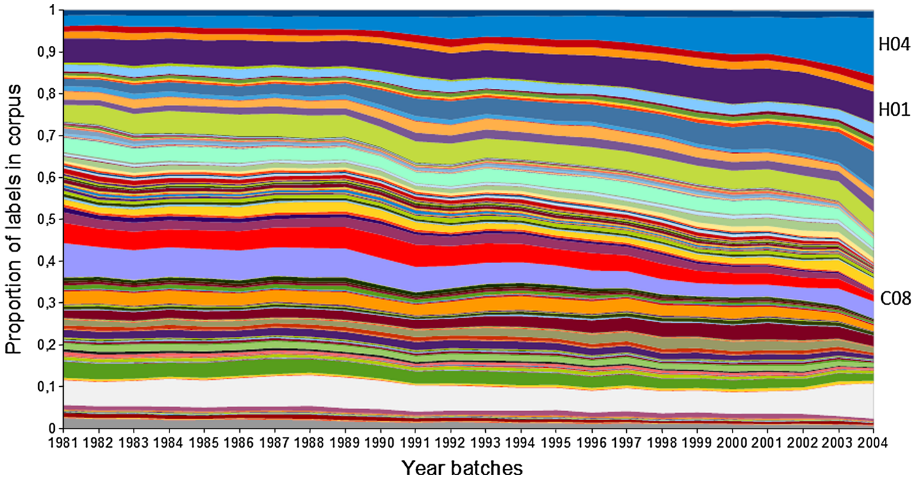

Fig. 2 Category label distribution (proportions) of English abstracts in CLEF-IP 2011 corpus. Each band depicts a different category over time

4.1 Category distributions

First we examined the distributions of the category labels in subsequent years in the CLEF-IP 2011 corpus.[16] Figure 2 shows the proportions of label occurrences for the different categories over time.

The figure shows a gradual change of category sizes in the corpus. Category sizes do not change abruptly between consecutive years, but over a longer time period certain categories grow substantially, e.g. H04 *Electric Communication Techniques*, while others decline, e.g. C08 *Organic Macromolecular Compounds*. However, most categories remain more or less the same relative size over time. The size imbalance (discussed in Sect. 1) between the categories is also clearly visible: The majority of the 121 categories contain few documents. In this corpus there are no categories that first shrink and then revive at a later point in time, which would be an indication of recurring concept drift. Instead, once a category starts to shrink or grow, it continues to do so. Additional analysis showed that for those documents that have multiple labels, there are no substantial changes in label combinations over time in the corpus.

What does this entail for classification in this corpus? A gradual change means that—at least for sampling purposes—the differences in category distributions are fairly small between consecutive years. This suggests that it is safe to use documents which are a few years older than the incoming new documents to train classifiers.

4.2 Terms used within the categories

Second, we investigated the changes within the term sets per category. It appeared that, except for the smallest categories, where data sparseness prevented a detailed analysis, all

[16] Unlike the statistics shown in Sects. 4.2 and 4.3, Fig. 2 comprises data from the entire CLEF-IP 2011 (starting from 1981), rather than the subcorpus described above. However, the subcorpus was created through random sampling at the different time points, and consequently has the same underlying category label distribution.

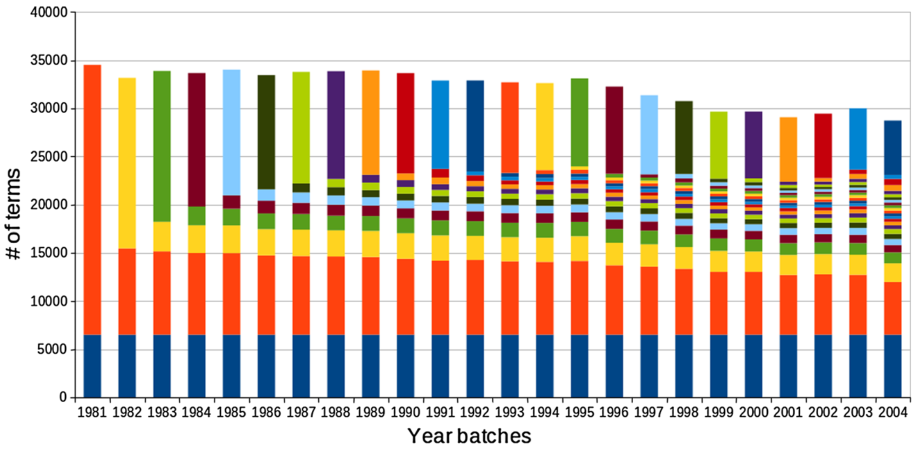

Fig. 3 Terminology shift in the H01 category in CLEF-IP 2011 subcorpus

categories showed similar emergence and fate of the terms (per year). We use the H01 *Basic Electric Elements* category to illustrate the phenomenon. We chose that category because it is large and stable in terms of size, and the pattern of emergence and fate of terms for this category is representative for the majority of the categories.

We took all terms occurring in the subcollection of documents for category H01 (not ranked in any way). For each term, we extracted from the corpus the year in which the term occurred for the first time for this category. Then we counted the number of terms originating per year. In Fig. 3, the differently coloured bands indicate the number of terms per year of the first appearance in the subcorpus. The bands show the number of terms originating in each year, but hold no information on the fate of individual terms. In other words, a term that occurs only in the 1981 and 1998 batches, will affect the height of the orange (second to lowest) band only in those years on the X-axis; the absence in the other years cannot be seen in this graph. The blue terms (lowest band) represents the number of 'stable' terms, that is, terms which appear in all years and remain in the corpus constantly.

Figure 3 reveals some interesting facts: First, a substantial amount of terms (on average 28 %, i.e. the 'stable' terms in the blue band) reappears in the term sets for each year. These terms are a mixture of (stable) category-specific terms and more general terms that occur frequently in technical documents. Second, each year around 20 % of the terms are newly introduced in the corpus. These terms can be seen in the top band for that year. This discovery is reminiscent of the category-specific terms discussed by Ma et al. (2009) (see Sect. 2). Only a small portion of these novel terms reappear in the subsequent year(s). The only exception to this observation is the broad orange band of terms introduced in 1981. These are the terms in the 1981 batch that do not re-appear in *all* subsequent years. One reason why this band is so broad is because 15,000 documents is a relatively small sample given the large number of categories in the subcorpus (121). With larger batches a larger proportion of the terms from 1981 band would have been present in all years, and considered as stable terms.

We also examined the changes in the term distribution in 10 smaller categories and some fast growing categories (A61; not shown here). In all cases we found an overall pattern similar to the one shown in Fig. 3. In small categories there tends to be a smaller number of 'stable' terms (blue band). Furthermore, while each year introduces novel terms, these are unlikely to re-appear in subsequent years. Both phenomena are clear indications of data sparseness. For growing categories such as A61, we find—unsurprisingly—a strong correlation between the number of new terms and category size. Moreover, there is a stronger tendency to inherit terms from previous years: New terms are added, rather than that they replace older ones.

What does this show us about concept drift in the patent corpus? For any term set in a given year, the majority of the terms have already been introduced into the corpus at an earlier stage, i.e. can be found in older training data. This also points to gradual drift. When selecting training data for the most up-to-date classifier, the overlap with the data distribution of previous years is relatively large, so that including older training documents will—likely—not harm classification accuracy. If the band of 'stable' terms (blue) were broader, the data distributions over time would be so similar that static sampling could be applied.

The quick disappearance of many newly introduced terms may have two causes: (a) most of these new terms are hapaxes, i.e. words that occur only once in the corpus and are thus of little importance; or (b) these terms are the 'fashionable' terms of that year and are evidence of concept drift. While (a) is certainly valid: 71 % of the novel terms in 2004 are hapaxes, the fact that some of these terms reappear in subsequent years reveals that a subset of these new terms are persistent additions to the language used to describe this category. All bands introduced after 1981 become thinner over time and do not increase at a later stage, which shows that there is no recurring concept drift in this corpus. These results suggest that it is beneficial to to update classifiers from time to time.

4.3 Category similarity

In this section we analyze the change of category similarities over time, which are caused by changes in the term distributions in the individual categories. For each category c in each year i, we created a subcorpus d by concatenating all documents with that category label c in that year i. We then created a term vector V_d which contains all terms that appear in this subcorpus, weighted by their TF-IDF weights. These weights are calculated as follows: (a) term frequency is the raw term frequency of the term t in the subcorpus $f_{(t,d)}$; (b) document frequency is the number of subcorpora in which term t appears, divided by the total number of subcorpora (D). TF-IDF weights are then calculated using formula 1.

$$tfidf_{t,d} = f_{(t,d)} * log_2\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right) \tag{1}$$

For each category, we then calculated the cosine similarity between the term vector of that category and the term vectors of the 120 other categories in the corpus, at the different time points in the corpus.

This yielded 121 × 120 time series, i.e. cosine similarity scores over time, for which we compute a linear fit function. If the gradient was not significantly different from zero, the categories did not become more (dis)similar over time. Given the very large number of tests, one would expect a substantial number of pairs for which the gradient would be non-zero at the $p = 0.05$ level. We were not able to find interesting patterns for particular categories, but we did find systematic behaviour related to category size, which is illustrated here for the H01 category *"Basic Electric Elements"*, which we also used to demonstrate the change in the use of terms.

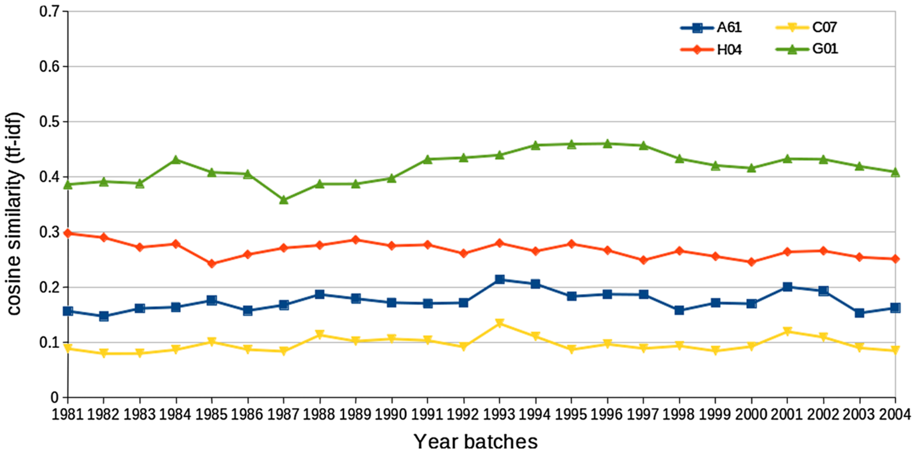

Fig. 4 Cosine similarity between H01 category and H04, G06, A61 and G01 categories in CLEF-IP 2011 subcorpus

Figure 4 shows the category similarities over time of the four largest categories in the corpus with the H01 category. As we can see, overall cosine similarity scores are rather low, which indicates that it is easy to distinguish between at least these four categories. Furthermore, the category similarities remain more or less constant over time. In other words, the larger categories do not become more or less similar to H01 over time. Interestingly, even A61, a category with grows over time, does not become more similar to H01.

Figure 5 shows the similarities between four intermediate categories and the H01 category. We can see that for the smaller categories, there is some change in category similarity over time. Some categories become more similar, others more dissimilar to the H01 category: Consider G11 *"Information Storage"*, which started out somewhat similar to H01, but becomes more dissimilar over time. In general, we found that there was a lower average category similarity between smaller categories and H01 compared to the larger categories described above. We assume that this is a consequence of the data sparseness in the smaller categories, which causes more internal variation (compared to the stable H01 category).

4.4 Summary: concept drift in the patent corpus

We can conclude that there is gradual concept drift in the patent corpus. Both in the class distributions and term distributions, we see no evidence of abrupt changes, but rather gradual shifts in category sizes and in the terms used to describe patents in a category. Change takes place over many years and is not recurring. However, it should be noted that more abrupt changes may exist at lower levels in the classification hierarchy. IPC classes themselves consists of different subclasses, groups and subgroups in some of which ground-breaking innovation that introduces new concepts or terminology may cause more abrupt changes. The category-internal changes do not lead to substantial changes in category similarity over time.

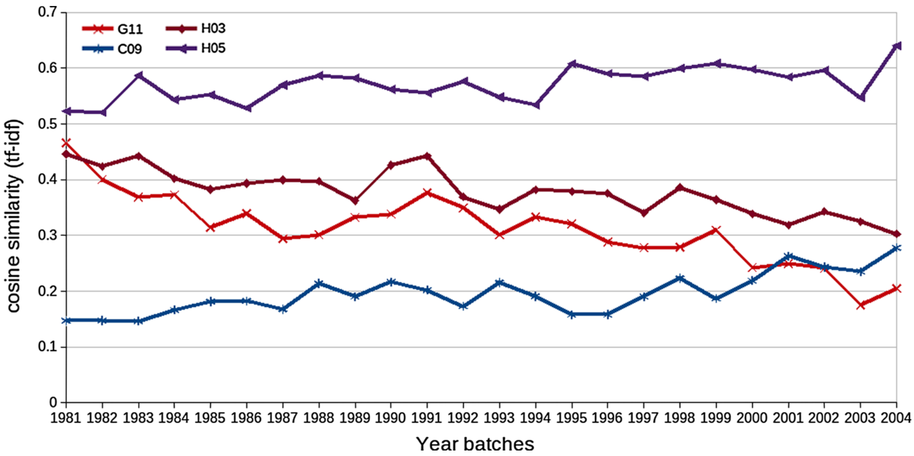

Fig. 5 Cosine similarity between H01 category and G11, C09, H03 and H05 categories in CLEF-IP 2011 subcorpus

5 Impact of time distance on classification performance

In the previous section we illustrated the gradual changes that occur over time in the data distribution(s) in the corpus. In this section we examine how these changes may cause mismatch between training set and test set distributions that might affect classification accuracy. For this purpose we examine the difference between static and temporally-aware sampling. We will discuss the impact of temporal variation for the larger and smaller categories, and examine the position of novel terms in the classifier models (class profiles[17]).

We examined the impact of (increasing) temporal distance through a series of experiments in which classifiers were trained on one year (15,000 documents), and then tested on all subsequent years in the corpus. In short, the training set remained constant, and we varied the test years. In these experiments there are multiple test sets, each comprising 15,000 documents, which may differ between themselves in difficulty. Therefore, we added a static sampling classifier as reference. This classifier was trained on 15,000 documents sampled randomly from all documents leading up to the test year. For example, the training material for the static sampling classifier used to test with the 1992 test set, was randomly selected from all available documents between 1981 and 1991. Please note that for the 1982 test set, the training sets of the baseline and the 1981 classifiers coincide.

Figure 6 shows the classification accuracy (F1, micro-averaged) of both the static sampling classifiers and the classifier trained on the documents from the year 1981 on the test sets consisting of the documents from each of the years 1982 until 2004. The figure shows that some test years are easier to classify than others. For example, for the 2003 and 2004 test set both classifiers achieve significantly higher scores than for the 1997 test set. Further analysis showed that this is a consequence of the relative growth of some of the large categories and the near disappearance of some smaller categories in the 2003 and 2004 batches (see also Sect. 4.1). Despite the variation between the test sets, we can clearly

[17] For each category, the Winnow algorithm outputs a set of the discriminating terms and their associated weights for that category.

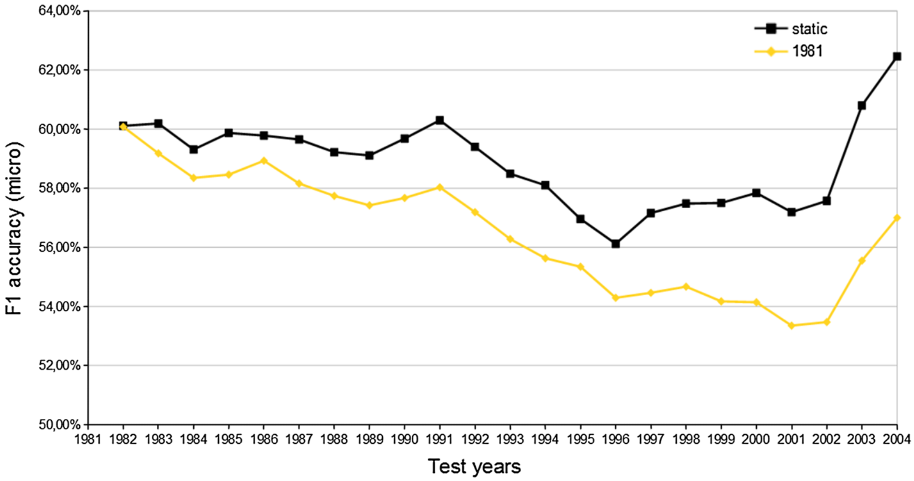

Fig. 6 *Black line* Classification performance on yearly batches with a classifier trained with 15,000 randomly selected documents. *Yellow line* Classification performance with a classifier trained with 15,000 documents from the year 1981 (F1 accuracy scores, micro-averaged) (Color figure online)

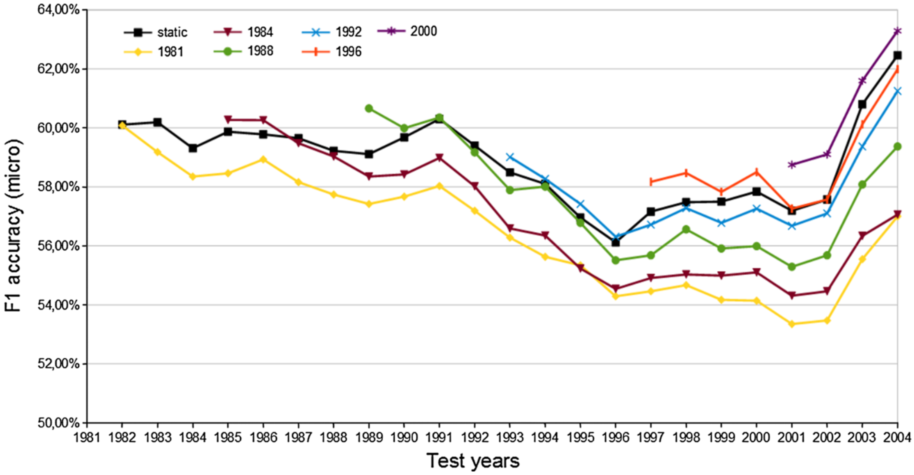

Fig. 7 Impact of recency for classifiers trained with documents from 1981, 1984, 1988, 1992, 1996 and 2000 (F1 accuracy scores, micro-averaged). (cf. caption of Fig. 6 for details.)

see a recency effect: The bigger the time distance between training and test set, the more classification accuracy drops. Interestingly, the classification accuracy of the 1981 classifier drops below the static sampling classifier score immediately. On the 1989 test set, the difference between classifiers becomes significant.

Figure 7 shows that classifiers trained on more recent data show similar behaviour as the classifier trained with the 1981 data: When the time distance between training and test set is small, these classifiers score better than the static sampling classifier. This effect is short-lived, however. All individual classifiers outperform the static classifier in the first couple of years, but then drop below the static classifier.

Table 2 Snapshots of the class profiles of the categories B60 (*Vehicles in general*) and H03 (*Basic electronic circuitry*), trained on 1981, 1991 and 2001 data, respectively

Rank	(a) Top ranking terms for B60 profile			(b) Top ranking terms for H03 profile		
	1981	1991	2001	1981	1991	2001
1	vehicle	vehicle	vehicle	successive	input	oscillator
2	tyre	tyre	tyre	gate	coding	interleave
3	brake	tread	*airbag*	transistor	radio	gain
4	rim	passenger	automobile	logic	*transistor*	*decode*
5	car	trailer	windscreen	encoder	conversion	*viterbi*
6	trailer	automobile	tread	cascade	value	filter
7	rear	car	roof	amplifier	stage	*bias*
8	wheel	wheel	car	tage	amplifier	digital
9	automotive	windscreen	pneumatic	input	cascade	circuit
10	elastomeric	roof	hybrid	radio	fall	error

Terms that are not in the 'stable' set are italicized

Analysis of the macro-averaged results (not shown here) showed similar effects of the time distance between the training and test data, although the recency effect lasts even shorter for the smaller categories. This means that the smaller categories are often so sparse that even though novel terms are introduced into the corpus each year, they occur too infrequently to aid classification performance. We can conclude that the effects of temporal variation are not clearly visible due to data sparseness in the small categories.

We also examined the effect of the novel terms in categories that are large enough to show a recency effect. Table 2 shows the top-ranking terms in the 1981, 1991 and 2001 class profiles of the B60 *"Vehicles in general"* and H03 *"Basic Electronic Circuitry"* categories. For these categories each more recent classification model (class profile) achieved consistently better results on the 2004 test set. By comparing them side by side we can get a snapshot of the change of individual terms over time. The table shows that most of the top ranking terms (ranked on Winnow weight) are the 'stable' terms that we discussed in Sect. 4.2. Exceptions are designated in italics. These terms were introduced in the late eighties, early nineties. Analysis of the most recent terms, i.e. terms that are introduced in the last five years, show that these are typically situated at much lower ranks (>1,000) in the class profiles.

6 Recency versus training window size

In the previous section we found that classifiers trained on the most recent data achieve the highest classification accuracy. It should be noted, however, that training samples of 15,000 documents are relatively small, considering the large number of categories (121) in the subcorpus. Especially for the smaller categories too little training material is available to adequately capture time effects. We expect that more data will give better results. In order to get better classification results, we should therefore extend the training sample, i.e. the window size over the training data, so that potentially informative terms become frequent enough in the training set to get through initial term selection and actively play a

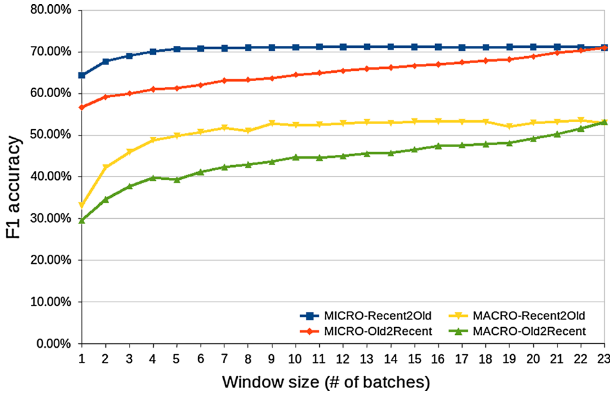

Fig. 8 Training sample size and recency effect of training data on 2004 test set. From left to right the training sets are increased by 15,000 documents per step. In Recent2Old additional documents become older at each step. In Old2Recent additional documents become more recent. At step 23 the training sets of Old2Recent and Recent2Old are identical

role during training. By extending the training window we also inevitably introduce older terms, which might be irrelevant or even cause nuisance, into the data distribution.

We performed a series of experiments in which we increased the number of documents in the training data with one year, i.e. 15,000 documents, for each new classifier, while the test set (the 2004 batch) remained constant. We did this in two ways. In the Recent2Old condition, we increased the number of training documents in the original classifier trained with data from the year 2003 by successively adding data from previous years to the training set. In the Old2Recent condition we started with the classifier trained with data from 1981, and successively added data from later years. Figure 8 shows the classification accuracy results of the two sets of classifiers on the 2004 test set. The figure contains both the micro- and macro-averaged values.

Figure 8 shows that for the same amount of training data, i.e. the same number of batches, a classifier trained on more recent documents significantly[18] outperforms a classifier trained on older data. In other words, given enough training material a recency effect can be seen in the smaller categories as well. In general we can conclude that adding more data improves classification accuracy, but starting with the most recent batch will lead to ceiling performance faster.

In case of the micro-averaged scores the classification accuracy of the Recent2Old run stabilizes fairly quickly (at around 5 batches, i.e. 75,000 documents). After this point, we can find no significant improvement in classification scores when adding more data. Perhaps equally interestingly, we do not see a negative effect of adding potentially irrelevant data to the training set. This may be thanks to the mistake-driven training strategy in the Winnow classifier.

When looking at the macro-averaged values, i.e. non-weighted average of the category scores, we see a slower accuracy increase as the size of the training set grows, but the same

[18] For the macro-averaged results, we used Wilcoxon paired rank tests to compare the improvement of individual category accuracy scores (F1) in the two runs. There was significant improvement ($p < 0.05$) of the Recent2Old scores compared to the Old2Recent scores, for all data points except those of classifiers at step 23.

stabilization, this time at around 10 batches (150,000 documents). This is the amount of documents needed to fully populate the feature space for the smaller categories.

We can conclude that when selecting data to train patent classifiers, the most recent data is the best. However, for an mistake-driven classifier such as Winnow adding older data does not have a negative effect. At around 150,000 training documents, the macro-averaged scores show that the performance of even the classifiers of the smaller categories stabilizes.

7 Impact of text representations

In previous research D'hondt et al. (2012, 2013) we explored the use of different text representations to capture relevant information in patent texts. We found that, given enough training data, adding certain phrasal features—more specifically PoS-filtered 2-skip-2-grams—significantly improves classification accuracy over unigram-only runs. However, the training material in those experiments was acquired through static sampling. In this section we investigate this effect in the light of concept drift.

Phrasal features are much sparser than words, and consequently have a less well-defined distribution. Given the changes of word features over time reported in Sect. 4.2, it is quite possible that temporal distance will have a larger effect on phrasal features. Still, in the previous section we have seen that the smaller categories in the corpus suffer from data sparseness, and the recency effect is much more apparent in the larger categories. Adding more (distinguishing) features to the feature space might help smaller categories to achieve higher accuracy over time.

We conducted a set of experiments in which we compared classifiers trained on words only and on words combined with skipgrams. The choice of training and test material is

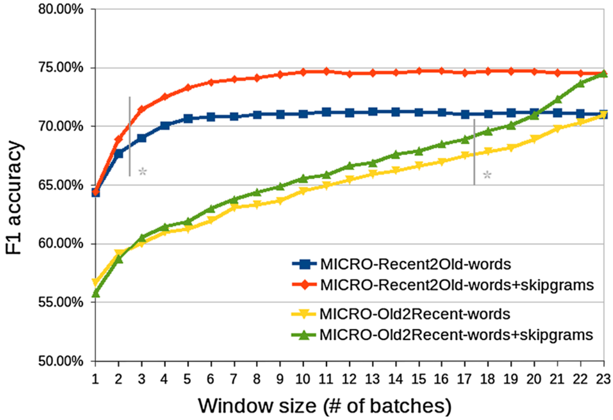

Fig. 9 Micro-averaged F1 scores show the impact of combining text representations (words-only vs words + skipgrams) and the recency effect of training data on the 2004 test set. From *left* to *right*, training sets are increased by 15,000 documents per step, either with older (Recent2Old) or newer (Old2Recent) data. Significance of differences between results from the words-only and words + skipgrams runs is calculated with ranges for 95 % confidence intervals, both for the Recent2Old and Old2Recent data selection. The *grey line* indicates the training window size where results from two classifiers become significantly different. *Left* of *grey line*, differences are insignificant; *right* of *grey line*, differences are significant

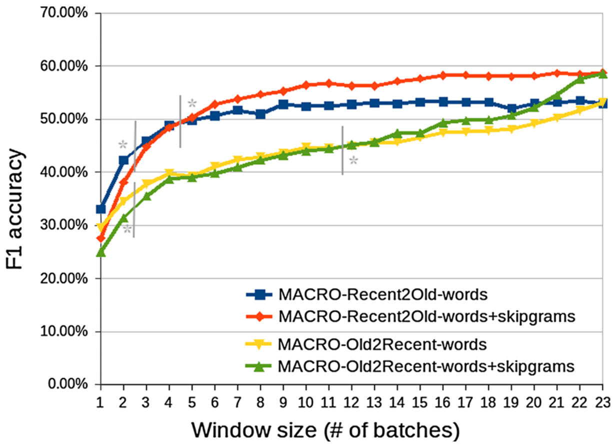

Fig. 10 Macro-averaged F1 scores which show the impact of combining text representations (words-only vs words + skipgrams) and the recency effect of training data on the 2004 test set. Significance of differences between results is calculated with Wilcoxon signed-rank test ($p < 0.05, n = 121$). The *grey lines* indicates sample sizes where differences between the runs become (in)significant (cf. caption of Fig. 9 for details)

similar to the procedure described in Sect. 6. Figures 9 and 10 show the interaction between text representations (words-only versus words + skipgrams) and recency on classification accuracy, for the micro-averaged and macro-averaged F1 scores, respectively.

Figure 9 shows clearly that—for the micro-averaged scores, which are dominated by the large categories—adding phrases to the words in the most recent documents quickly (after 45,000 documents) leads to significant improvements in classification accuracy over the words-only run. In the Old2Recent runs, significant improvement of the words + skipgrams representation over the words-only representation is only achieved with a window size of 18 batches (270,000 documents). This suggests that only the most recent phrases have a positive impact on classification accuracy, while phrases in older batches have no (significant) impact.

We find a similar, though less pronounced, effect for the phrasal features when considering the macro-averaged accuracy scores in Fig. 10. For smaller categories adding more recent phrases leads to significant improvements, although more training data is needed. Interestingly, for the first two windows (n = 1 and n = 2) the words + skipgrams runs performed significantly worse than their words-only counterparts. We suspect that with so little training material, no term selection took place and all terms were used for training. Consequently, vague and noisy phrasal terms like *have_be*, which normally do not make it through term selection, may have deteriorated classification accuracy.

We can conclude that for phrases—as for words—more training data is always better, and starting from the most recent data gives the best results. Temporal variation has a stronger effect on phrasal features, because of their relative sparseness compared to words. Given enough data, both the performance of smaller and larger categories can be improved by adding phrasal features.

8 Conclusion

In this article we quantified the existence of concept drift in a large patent corpus and we investigated its impact on classification accuracy. We will now summarize our findings and consider their practical implications for building patent classification systems that can adequately deal with novel patent applications.

In the CLEF-IP 2011 patent corpus, we found evidence of slow but continuous changes (gradual drift) over time in the data distribution of all categories in the corpus. This drift is manifested in three ways: (a) The relative proportions of the categories change over time; (b) within the categories the feature distribution changes over time as new words are added each year; and (c) the (dis)similarities between the different categories in the corpus are not systematically affected by the category-internal changes.

Ignoring the temporal variation affects patent classification accuracy: We found that classification models built on older data quickly become less powerful as the temporal distance between training set and test set increase. When comparing temporally aware sampling with static sampling, classifiers trained on the most recent data outperformed the static sampling classifiers. However, we also found that the addition of older data (which may contain irrelevant terms) is not harmful to classification accuracy.

How does this translate to the practice of building an effective classifier for handling incoming patent applications?

- More training data will result in higher classification accuracy, as long as you start with the most recent data and go back in time. Note however that the patent corpus is highly imbalanced. Consequently, smaller categories need more data (and therefore a wider time window) to obtain good models.
- This approach might depend on the classification algorithm: The mistake-driven Winnow algorithm is very robust against data imbalance and can deal with the noise created by adding older (and perhaps irrelevant) terms in the training data set.
- The concept drift in the patent corpus is monotonic and not recurring: Consequently, older classification models become obsolete and can be discarded.
- Drift is very gradual. Unlike a news corpus in which ever-changing content prompts the need for an online classifier that is updated on a daily basis, patent classifiers can easily be trained in batches covering longer periods of time.
- Classification accuracy can be improved by combining unigram terms with more informative text representations. We confirmed the finding in previous work that PoS-filtered 2-skip-2-grams adequately capture Multi-Word Terms which are useful classification features. While the effect of adding phrasal features to unigrams is short-lived—since phrasal feature are more inherently more sparse, they are more easily affected by temporal variation—adding phrasal features from the most recent training data can significantly improve classification performance.

In the current article we have focussed exclusively on the (relative) re-usability of content features throughout changing feature distributions. However, if one were to go beyond that, two promising avenues of research could be identified which we offer up as suggestions to the community for further research:

First, the problem of dealing with unseen terms in incoming patents: In Sect. 4.2, we found that each year on average around 20 % novel terms are introduced in the corpus. Although many of these terms are hapaxes and disappear quickly, some remain and become part of the vocabulary that characterizes a category. If these terms could be integrated into the classification models, they might improve classification performance.

But since newly introduced terms were absent from the training they cannot make their contributions. We envision an approach in which the incoming test documents are split into documents with seen and unseen vocabulary and use classification results obtained with the documents with seen vocabulary to build models that include the unseen terms. Even if the number of new terms is small, combining the results of classification of documents with seen and unseen vocabulary might improve overall classification.

Second, in this paper we only used the full-text of the abstracts in patent documents. However, patent documents come with rich metadata, i.e. information on the context of the invention. This information is fully grounded in time. While the use of metadata is not without problems for patent classification (Richter and MacFarlane 2005), it would be worthwhile to investigate how information on the changes of assignee, inventor, patent examiner, etc. may aid patent classification in a time context. It should be mentioned, however, that the most potentially effective metadata are very sparse and are therefore probably best used in a separate classifier.

We expect that the insights formulated in this article will be of interest to the patent community as a whole, and may serve as the starting point for future research on this topic.

References

Benzineb, K., & Guyot, J. (2011). Automated patent classification. In M. Lupu, K. Mayer, J. Tait, & A. J. Trippe (Eds.), *Current challenges in patent information retrieval* (Vol. 29, pp. 239–261). Berlin: Springer.

Carmona-Cejudo, J. M., Baena-García, M., Bueno, R. M., Gama, J., & Bifet, A. (2011). Using gnusmail to compare data stream mining methods for on-line email classification. *Journal of Machine Learning Research-Proceedings Track*, *17*, 12–18.

Cohen, A., Bhupatiraju, R., & Hersh, W. (2004). Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage. In *Proceedings of the thirteenth text retrieval conference-TREC*.

Dagan, I., Karov, Y., Roth, D. (1997). Mistake-driven learning in text categorization. In *Proceedings of 2nd conference on empirical methods in NLP*, Providence, pp. 55–63.

D'hondt, E., Verberne, S., Weber, N., Koster, K., & Boves, L. (2012). Using skipgrams and pos-based feature selection for patent classification. *Computational Linguistics in the Netherlands Journal*, *2*, 52–70.

D'hondt, E., Verberne, S., Koster, C., & Boves, L. (2013). Text representations for patent classification. *Computational Linguistics*, *39*(3), 755–775.

Fawcett, T. (2003). "In vivo" spam filtering: A challenge problem for KDD. *ACM SIGKDD Explorations Newsletter*, *5*(2), 140–148.

Forman, G. (2004). A pitfall and solution in multi-class feature selection for text classification. In *Proceedings of the twenty-first international conference on machine learning, ICML '04* (pp. 38–45). New York, NY: ACM.

Forman, G. (2006). Tackling concept drift by temporal inductive transfer. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '06* (pp. 252–259). New York, NY: ACM.

Frantzi, K., Ananiadou, S., & Tsujii, J. (1998). The C-value/NC-value method of automatic recognition for multi-word terms. In *Proceedings of the second European conference on research and advanced technology for digital libraries, ECDL '98* (pp. 585–604). London: Springer.

Galavotti, L., Sebastiani, F., & Simi, M. (2000). Experiments on the use of feature selection and negative evidence in automated text categorization. In *Proceedings of research and advanced technology for digital libraries, 4th European conference*, Lisbon, pp. 59–68.

Ja, Gama, Medas, P., Castillo, G., & Rodrigues, P. (2004). Learning with drift detection. In A. Bazzan & S. Labidi (Eds.), *Advances in artificial intelligence SBIA 2004, lecture notes in computer science* (Vol. 3171, pp. 286–295). Berlin: Springer.

Joachims, T. (1999). Making large-scale support vector machine learning practical. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in Kernel methods* (pp. 169–184). Cambridge: MIT Press.

Kelly, M., Hand, D., & Adams, N. (1999). The impact of changing populations on classifier performance. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '99* (pp. 367–371). New York, NY: ACM.

Klimt, B., & Yang, Y. (2004) The enron corpus: A new dataset for email classification research. In *Proceedings of the 15th European conference on machine learning, ECML 2004*, Vol. 15, p. 217. Berlin: Springer.

Klinkenberg, R. (2004). Learning drifting concepts: Example selection vs. example weighting. *Intelligent Data Analysis*, *8*(3), 281–300.

Koster, C., & Beney, J. (2009). Phrase-based document categorization revisited. In *Proceedings of the 2nd international workshop on patent information retrieval, PaIR '09* (pp. 49–56). New York, NY: ACM.

Koster, C., & Seutter, M., Beney, J. (2003). Multi-classification of patent applications with winnow. In M. Broy, A. V. Zamulin (Eds,). *Ershov memorial conference, Lecture Notes in Computer Science*, Vol. 2890 (pp. 546–555). Berlin: Springer.

Koster, C., Beney, J., Verberne, S., & Vogel, M. (2011). Phrase-based document categorization. In M. Lupu, K. Mayer, J. Tait, & A. J. Trippe (Eds.), *Current Challenges in Patent Information Retrieval* (Vol. 29, pp. 263–286). Berlin: Springer.

Koychev, I. (2000). Gradual forgetting for adaptation to concept drift. In *Proceedings of ECAI 2000 workshop on current issues in Spatio-Temporal reasoning*.

Kuncheva, L. (2004). Classifier ensembles for changing environments. In F. Roli, J. Kittler, & T. Windeatt (Eds.), *Multiple classifier systems, lecture notes in computer science* (Vol. 3077, pp. 1–15). Berlin: Springer.

Lebanon, G., & Zhao, Y. (2008). Local likelihood modeling of temporal text streams. *Proceedings of the 25th international conference on Machine learning—ICML '08* (pp. 552–559). New York, NY: ACM Press.

Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, *5*, 361–397.

Liu, R., & Lu, Y. (2002). Incremental context mining for adaptive document classification. In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 599–604). New York: ACM.

Ma, C., Lu, B. L., & Utiyama, M. (2009). Incorporating prior knowledge into task decomposition for large-scale patent classification. In W. Yu, H. He, & N. Zhang (Eds.), *Advances in neural networks ISNN 2009, lecture notes in computer science* (Vol. 5552, pp. 784–793). Berlin: Springer.

Mourão, F., Rocha, L., Araújo, R., Couto, T., Gonçalves, M., & Meira, W. J. (2008). Understanding temporal aspects in document classification. In *Proceedings of the 2008 international conference on web search and data mining (WSDM '08)* (pp. 159–170). New York: ACM.

Nanba, H., Fujii, A., Iwayama, M., & Hashimoto, T. (2008). Overview of the patent mining task at the NTCIR-7 workshop. In *Proceedings of NTCIR-7 workshop meeting*, pp. 325–332.

Nanba, H., Fujii, A., Iwayama, M., & Hashimoto, T. (2010). Overview of the patent mining task at the NTCIR-8 workshop. In *Proceedings of NTCIR-7 workshop meeting*, pp. 293–302.

Oostdijk, N., Verberne, S.,&Koster, C. (2010). Constructing a broad-coverage lexicon for text mining in the patent domain. In Chair NCC, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*, Valletta, Malta.

Richter, G., & MacFarlane, A. (2005). The impact of metadata on the accuracy of automated patent classification. *World Patent Information*, *27*(1), 13–26.

Rocha, L., Mourão, F., Mota, H., Salles, T., Gonçalves, M. A., & Meira, W, Jr. (2012). Temporal contexts: Effective text classification in evolving document collections. *Information Systems*, *38*(3), 388–409.

Salles, T., Rocha, L., Pappa, G.L., Mourão, F., Meira, W. Jr, & Gonçalves, M. (2010). Temporally-aware algorithms for document classification. In *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval, SIGIR '10* (pp. 307–314). New York, NY: ACM.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing Management*, *24*(5), 513–523.

SanJuan, E., Dowdall, J., Ibekwe-SanJuan, F., & Rinaldi, F. (2005). A symbolic approach to automatic multiword term structuring. *Computer Speech and Language*, *19*(4), 524–542.

Schlimmer, J., & Granger, R, Jr. (1986). Incremental learning from noisy data. *Machine Learning*, *1*, 317–354.

Scholz, M., & Klinkenberg, R. (2007). Boosting classifiers for drifting concepts. *Intelligent Data Analysis*, *11*(1), 3–28.

Segal, R., & Kephart, J. (1999). Mailcat: An intelligent assistant for organizing e-mail. In *Proceedings of the third annual conference on autonomous agents* (pp. 276–282). New York, NY: ACM.

Šilić, A., & Dalbelo Bašić, B. (2012). Exploring classification concept drift on a large news text corpus. In *Computational linguistics and intelligent text processing*, pp. 428–437.

Tsymbal, A. (2004). *The problem of concept drift: Definitions and related work. Tech. Rep. TCD-CS-2004-15*, Computer Science Department, Trinity College Dublin.

van Halteren, H. (2000). The detection of inconsistency in manually tagged text. In Proceedings of LINC-00.

Verberne, S., Vogel, M., & D'hondt, E. (2010). Patent classification experiments with the linguistic classification system LCS. In *Proceedings of the conference on multilingual and multimodal information access evaluation (CLEF 2010)*, Padua.

Žliobaitė, I. (2009). *Learning under concept drift: An overview*. Tech. rep.: Vilnius University.