Concept Profiles for Filtering Parliamentary Documents

Francisco J. Ribadas[1], Luis M. de Campos[2], Juan M. Fernández-Luna[2] and Juan F. Huete[2]
[1]*Departamento de Informática, E.S. Enxeñería Informática, Edificio Politécnico, Universidade de Vigo, 32004 Ourense, Spain*
[2]*Departamento de Ciencias de la Computación e Inteligencia Artificial, ETSI Informática y de Telecomunicación, CITIC-UGR, Universidad de Granada, 18071 Granada, Spain*

Keywords: Information Filtering, Parliamentary Documents, Text Classification, User Profiles.

Abstract: Content-based recommender/filtering systems help to appropriately distribute information among the individuals or organizations that could consider it of interest. In this paper we describe a filtering system to deal with the problem of assigning documents to members of the parliament potentially interested on them. The proposed approach exploits subjects taken from a conceptual thesaurus to create the user profiles and to describe the documents to be filtered. The assignment of subjects to documents is modeled as a multilabel classification problem. Experiments with a real parliamentary corpus are reported, evaluating several methods to assign conceptual subjects to documents and to match those sets of subjects with user profiles.

1 INTRODUCTION

The application of Information and Communications Technologies (ICT) has dramatically increased the capabilities of individuals, organizations and companies for appropriately storing, managing and retrieving their information. Centering on the so-called e-government (application of ICT in public administration), we are going to further focus in a parliamentary context. Parliaments, like other organizations, generate and also receive lots of information, which must be appropriately distributed among the people or institutions that could consider it of interest.

Particularly, let us consider a new document (either external or internal to the parliament) that must be distributed among the Members of the Parliament (MP). We would like to build a system that could help to determine which MPs are going to receive it. This decision should be based both on the content itself of the document and the specific interests of the MPs.

Therefore, this research falls within the field of content-based recommender/filtering systems (Belkin and Croft, 1992; Hanani et al., 2001; Lops et al., 2011; Pazzani and Billsus, 2007), which are systems that recommend an item (a document in our case) to a user based on a description of the item and a profile (Gauch et al., 2007) of the user's interests.

In order to build the MP profiles, we must learn in some way about the interests of each MP. We are going to automatically extract this information from the activities carried out by each MP within the parliament. Later, we must build some kind of model able to match the content of the document to be filtered against the MP profiles, in order to recommend the document to the top ranked MPs.

The activity in any parliament revolves around the concept of parliamentary initiative, whereby an action taken by an MP or political party is discussed in a plenary or specific area committee session. The transcriptions of all the MP speeches within these debates are collected in the form of records of proceedings.

In our case study[1], which is the Parliament of Andalusia (a region at Spain), parliamentary initiatives are tagged with one or more subjects or descriptors extracted from the EUROVOC thesaurus[2], which are manually assigned by parliamentary documentalists as being the best representation of its content. The main usefulness of these subjects is to allow faceted retrieval, in such a way that any user can easily retrieve the initiatives dealing with a given set of subjects. Our proposal is to use these subjects in another way: the subjects associated to the initiatives where a given MP takes part will be used to build his/her corresponding profile. Thus, our profiles will consist of lists of weighted subjects.

[1]And also in other cases, as the European Parliament.
[2]http://eurovoc.europa.eu/

Ribadas, F., Campos, L., Fernández-Luna, J. and Huete, J..
Concept Profiles for Filtering Parliamentary Documents.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR*, pages 409-416
ISBN: 978-989-758-158-8

But, if the MP profiles are composed of subjects, in order to match the document to be filtered with these profiles, it must also be represented by a list of weighted subjects. Our proposal is to train a classifier, using the content (the textual transcriptions of the speeches of the MPs) of the initiatives as attributes and the subjects as the categories or labels. As an initiative can be labeled with more than one subject, this is a multilabel classification problem. Given a new document to be filtered, we shall use the classifier to obtain a list of the more appropriate subjects for labeling it. Next, this (weighted) list of subjects will be matched against the MP profiles, in order to decide which are the MPs that could be more interested in the document.

The remaining of this paper is organized as follows: in Section 2 we describe the methods for user profile construction. Section 3 explains how to transform the document to be filtered into a set of subjects that will be matched against the user profiles. Section 4 describes the matching functions and the procedure employed to assign documents to their corresponding MPs. Finally in Section 5 we report the results of our experiments and in Section 6 we review the main contributions of this work.

2 USER PROFILE CONSTRUCTION

As we have already mentioned, we want to represent the interests and preferences of the MPs by means of user profiles. The two most important steps in the user profile building process are the acquisition of user information and the user profile representation (Gauch et al., 2007).

Regarding the acquisition of user information, it may be performed explicit or implicitly. In our case (and also in the general case), users are reluctant to provide their personal information and to fill questionnaires, so that an explicit approach is not feasible. Fortunately we have a source of public information about the MPs, in order to apply an implicit approach: the transcriptions of their speeches when discussing each initiative, within the parliamentary debates, and the subjects of the EUROVOC thesaurus associated to each initiative by the parliament documentalists.

Regarding user profile representation, the three main approaches are a set of weighted keywords, semantic networks, and a set of weighted concepts (Gauch et al., 2007). Weighted keywords is the most common user profile representation and the easiest to build: they may be automatically learned from documents relevant to the user (or directly given by the user). Semantic networks, where each node represents a concept, are more difficult to build and manage. They also must learn the terms associated with each network node, sometimes derived from ontologies or external knowledge resources such as ODP or WordNet. Weighted concepts are similar to the semantic networks, since they also have conceptual nodes (and sometimes relations between them), but in this case the nodes represent abstract topics of interest for the user instead of terms. They are also similar to the weighted keywords profiles, since they are usually represented as vectors of weighted concepts. In contrast to semantic networks, weighted concept profiles are trained on example texts for each concept a priori, and thus there exist relationships between vocabulary and concepts from the beginning. Thereby, the built user profiles can be more robust to variations in terminology.

In our case, as each initiative is already (reliabily) tagged with EUROVOC subjects, it is very easy to build a weighted concept profile based on these subjects. Therefore, this is the approach we shall follow to represent MP profiles. Nevertheless, we do not discard for future work to experiment with weighted keywords profiles (extracting the terms directly from the transcriptions of the speeches of each MP).

We are going to use two simple methods to build the MP weighted subjects profiles. In both cases, for each MP, we collect all the initiatives where he/she participates and extract the list of subjects assigned to these initiatives. In the first method, called SF (from subject frequency), the weight of each subject s is simply the number of initiatives where the MP participated which are labeled with this subject, $\mathrm{tf}_s(MP)$. In the second method we compute a kind of inverse document frequency of each subject s, idf_s, measured as $\log(\frac{N}{n_s})$. N is the number of MPs and n_s is the number of MPs who participated in at least one initiative labeled with s. The weight of each subject s for an MP is then $\mathrm{tf}_s(MP) \times \mathrm{idf}_s$. We call this method SFIDF.

3 ASSIGNING SUBJECTS TO DOCUMENTS

Given a document that must be distributed among the MPs, we must transform this input into a set of weighted subjects, in order to match this list against the user profiles. Our proposal is to use a multilabel text classifier to assign to this document a set of the most appropriated subjects (together with their weights). As depicted in Figure 1, the training data for building this classifier will be the content of the initiatives (the textual transcriptions of the speeches

of the MPs), as well as the associated subjects.

In the literature there are different approaches to deal with multilabel classification problems (Tsoumakas et al., 2010). There are simple approaches like Binary Relevance (BR) which does not consider dependencies between labels, or approaches like Label Powerset (LP) where every possible combination of labels is treated as a metalabel in order to train the classifier.

In our case, the large number of subjects in the EUROVOC thesaurus makes a LP approach unfeasible, so we have implemented a multilabel categorization scheme based on Classifier Chains (CC) (Read et al., 2011). This method offers a compromise between the simplicity of BR and the computational cost of LP, still being able to exploit label dependencies. CC is an ensemble based approach that iteratively builds a set of binary classifiers (one for each label), adding at each step a new feature that encodes the class predicted by the immediately previous classifier. Any binary classification algorithm can be employed to build those elementary classifiers. In our experiments we have evaluated the use of Support Vector Machines (SVM) as elementary classifiers, using the LibSVM (Chang and Lin, 2011) implementation, and also the use of Naive Bayes classifiers, using the implementation from the Weka machine learning environment (Hall et al., 2009). Additionally, unlike in the original definition of CC, the sequence of elementary classifiers follows the usage frequency of EUROVOC subjects in our dataset, starting with the most frequent ones instead of using a random ordering.

Since there are hierarchical relationships between EUROVOC thesaurus subjects, an alternative approach could have been modeling the subject assignment task as a hierarchical classification problem. Some improvements in categorization performance have been reported for several small and medium scale problems using the hierarchical top-down method named *Local Classifier Per Node Approach* in the taxonomy of hierarchical classification approaches presented in (Silla Jr. et al., 2011). However, the relatively small size of our initiatives collection with respect to the number of labels in the EUROVOC thesaurus does not advise applying this kind of approaches in our case. This is a consequence of a well known weakness of this kind of methods related with the poor performance of the local classifiers associated with infrequent or never used labels. In our corpus, classifiers created for the infrequent EUROVOC labels would have been trained with no enough positive examples and it is expected that they will provide us with inconsistent results.

We do have evaluated a similar but much simpler alternative taking advantage of the concept of microthesaurus used in EUROVOC. The first level of the label hierarchy arranged by EUROVOC thesaurus conforms a set of 96 microthesauri, which, with few exceptions, can be considered relatively independent. Thus, we have transformed the EUROVOC subject assignment task in a two phases multilabel classification problem: the first phase will attempt to predict the set of relevant microthesauri for a given document, and in the second phase the final set of subjects from each selected microthesaurus will be created using its corresponding local multilabel classifier.

4 PROFILE MATCHING

Once the document to be filtered has been classified to obtain its associated subjects, we have as the output a weighted list of subjects that try to describe its content. We call this list the document profile, as shown in Figure 2. The remaining step is to match this document profile against the MP profiles. To do that we have to define a matching function.

Several alternatives when addressing the problem of matching weighted concepts vectors have been proposed. In our experiments we have evaluated the suitability of the following set of matching functions, where $P1$ and $P2$ are two weighted vector profiles being matched and P_k is the weight of the k-th subject in the weighted vector depicting profile P.

Simple Matching. This measure simply computes de inner product between the two vectors representing the profiles being matched.

$$sim_{\text{SIMPLE}}(P1,P2) = \sum_{k=1}^{n}(P1_k \times P2_k)$$

Only subjects with non-zero weights in both profiles will contribute to the final score.

Cosine Similarity. This measure computes the cosine between the two weighted vectors representing the profiles being matched.

$$sim_{\text{COSINE}}(P1,P2) = \frac{\sum_{k=1}^{n}(P1_k \times P2_k)}{\sqrt{\sum_{k=1}^{n} P1_k^2} \times \sqrt{\sum_{k=1}^{n} P2_k^2}}$$

Inverse Euclidean Distance. This measure computes the euclidean distance between the two weighted vectors representing the profiles being matched. In order to interpret this distance as a similarity measure, it is inverted.

$$sim_{\text{EUCLIDEAN}}(P1,P2) = -\sqrt{\sum_{k=1}^{n}(P'1_k - P'2_k)^2}$$

where P' represents the normalized version of the weighted vector P (dividing by the maximum weight).

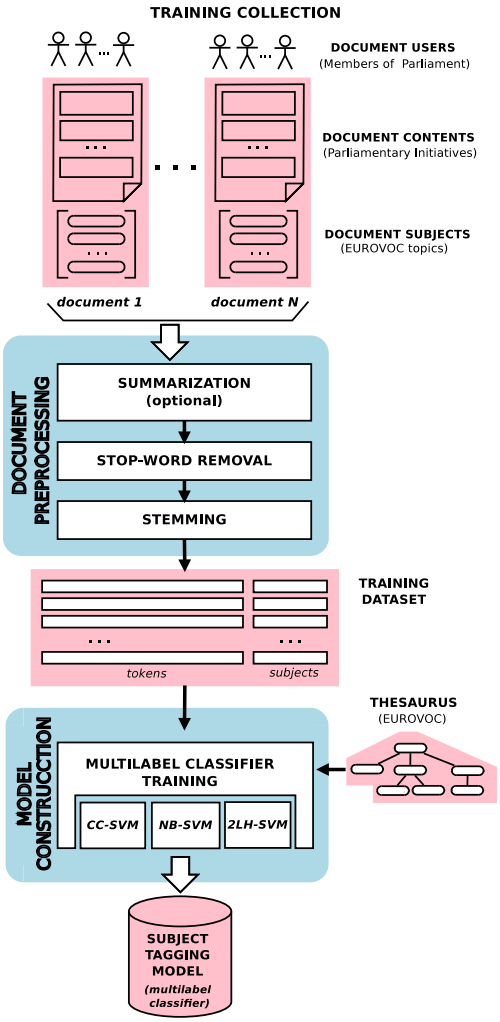

Figure 1: Training multilabel classifer.

Weighted Dice Coefficient. This measure computes a generalized version of the Dice coefficient over sets, where each element in the compared sets could have a non-zero weight.

$$sim_{\mathrm{DICE}}(P1,P2) = \frac{2 \times \sum_{k=1}^{n}(P1_k \times P2_k)}{\sum_{k=1}^{n} P1_k + \sum_{k=1}^{n} P2_k}$$

Document profiles are matched against all the available MP profiles using one of the previous functions to score that correspondence. As a result we obtain for each document an ordered list of candidate MPs according to the value of the matching function. The top n MPs with the highest scores will be assigned as relevant for the given document.

As is the case of multilabel classification, the fact that EUROVOC subjects are arranged in a tree leads us to try to take advantage of this additional hierarchical information. In our experiments we have assessed the usefulness of expanding the lists of subjects

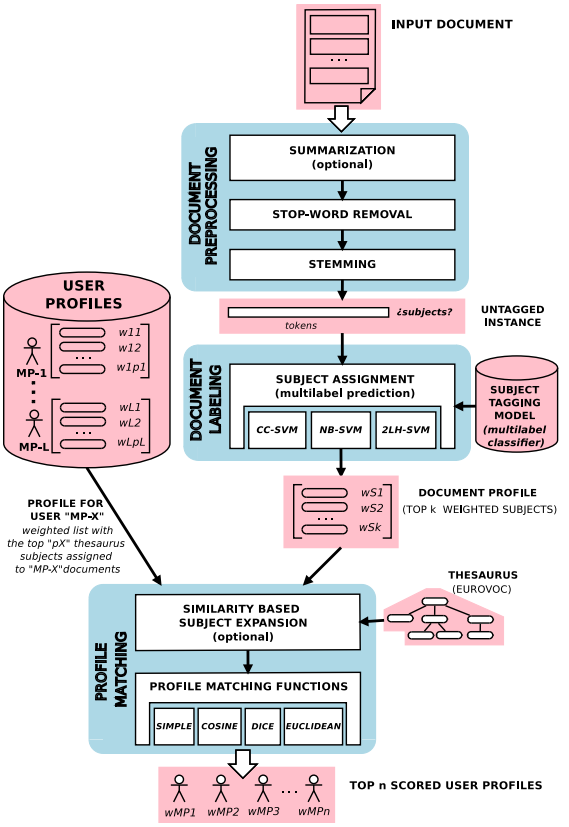

Figure 2: Profile assignment to input documents.

that describe both document and MPs profiles, adding similar subjects from the EUROVOC taxonomy. The intuition behind this scheme is that expanded profiles could provide a more generic and flexible description both for document and MP profiles.

Our proposal makes use of the semantic similarity measure on taxonomies defined in (Lin, 1998). This measure computes the information content for the involved concepts and for their lowest common ancestor. Those values are combined according to formula (1) to define a metric that quantifies the similarity between two concepts in the range [0,1]. In (1) s_i and s_j are concepts in a taxonomy, $LCA(s_i,s_j)$ represents their lowest common ancestor and $P(s_k)$ is an estimation of the probability assigned to concept s_k. In our case this probability is computed as the ratio between the number of EUROVOC subjects belonging to the subtree rooted at label s_k and the total number of subjects in EUROVOC thesaurus. Similarity values close to one correspond with similar concepts closely located in the taxonomy which are also expected to be semantically close, whereas 0 means total dissimilarity between two unrelated concepts.

$$sim(s_i,s_j) = \frac{2 \times logP(LCA(s_i,s_j))}{logP(s_i) + logP(s_j)} \qquad (1)$$

Table 1: Multilabel categorization and text preprocessing performance.

	CC-SVM			CC-NB			2LH-SVM		
	MiP	MiR	MiF	MiP	MiR	MiF	MiP	MiR	MiF
full text	0.3701	0.4712	0.4145	0.3462	0.4193	0.3792	0.3587	0.4686	0.4063
summary 75%	0.3612	0.4521	0.4015	0.3367	0.4067	0.3684	0.3502	0.4539	0.3953
summary 50%	0.3321	0.4297	0.3746	0.3202	0.3876	0.3506	0.3261	0.4313	0.3713
summary 25%	0.3121	0.3988	0.3501	0.2931	0.3657	0.3253	0.3079	0.4023	0.3488
manual summary	0.2081	0.3241	0.2534	0.1753	0.2701	0.2126	0.1923	0.3302	0.2430

For our subject expansion scheme we have followed the following strategy. For each subject belonging to a profile (both document profiles and MP profiles) a list of similar subjects within a neighborhood of 3 hops on the EUROVOC hierarchy is added to create the expanded profile. The weights of these neighbor subjects are calculated by multiplying the weight of each original subject by the Lin's similarity score between the original and each one of the neighbor subjects (and adding these weights if an added subject is neighbor of more than one original subject).

5 EXPERIMENTS

In order to validate our proposals, we have carried out an experimental study with a set of parliamentary documents. The document collection we use is the set of initiatives (5258) from the 8^{th} term of office of the Parliament of Andalusia at Spain[3], marked up in XML (de Campos et al., 2009). These initiatives contain a set of 12633 different interventions of the MPs (on the average there are 2.4 interventions per initiative). The initiatives have been tagged by parliamentary documentalists using subjects from the EUROVOC thesaurus. From this set we shall build the profiles only for those MPs who participate in more than 10 different initiatives, giving a total of 132 MP profiles. The average number of interventions per MP is 92.5 (with a standard deviation equals to 71.3). From this initial corpus we removed those initiatives that are labeled with no subject. The result is a final set of 4523 initiatives, with an average of 3.89 subjects per initiative.

We have used the repeated holdout method (Lantz, 2013) to evaluate the behavior of our system: the set of initiatives is randomly partitioned into a training set (80%) and a test set (20%), and we repeat this process five times. The reported results are the averages over the five different rounds. The training set is used to learn the MP profiles (using the subjects associated to the initiatives) and also to train the classifier (using the textual content of the initiatives, see Figure 1). The initiatives in the test set are used to classify them (using their textual content) and to obtain (through the matching of the proposed subjects with the profiles) the ranked list of MPs that are probably interested in them, as shown in Figure 2.

Obviously we use in this process neither the information related to the MPs who participate in each test initiative, nor its associated subjects. As the ground truth, we have considered that a test initiative is only relevant to those MPs who participate in it[4].

Our experiments have been structured in two phases. First, we have evaluated the performance of the various multilabel classification schemes described in Section 3 to deal with the characterization of initiative contents using EUROVOC subjects. In this first set of experiments we have also evaluated the effect of using different textual representations of initiative contents, like manually created extracts provided by parliament documentalists, the full text of the initiative transcripts or automatically generated summaries of these transcripts. The second phase of our experiments deals with the suitability of the profile matching procedure proposed in Section 4.

5.1 Text Processing and Categorization

Given the diversity in contents and structure shown by the initiatives included in our corpus, we have employed a very simple automatic summarization procedure in order to identify the most informative paragraphs able to characterize each initiative contents. We have followed a sentence extraction approach relying on conventional information retrieval tools. To do this every initiative in our test collection is split into their constitutive paragraphs. The textual content for each one of these paragraphs is stemmed using a Spanish stemmer from the Snowball project. A standard list of Spanish stop-words is employed to select the actual content holder terms. The surviving terms and overlapping bigrams of those selected index terms are indexed using the Apache Lucene indexing engine. This way, every document included in that index corresponds to one of the paragraphs included in the

[3]http://www.parlamentodeandalucia.es

[4]This is a rather conservative assumption, because it is quite reasonable to think that an initiative can also be relevant to other MPs.

Table 2: Profile matching performance using SF profiles.

	ideal			ideal with expansion			CC-SVM			CC-SVM with expansion		
	MAP	RPrec	R@10	MAP	RPrec	R@10	MAP	RPrec	R@10	MAP	RPrec	R@10
simple	**0.5173** †	**0.3419** †	**0.6876** †	**0.4955** †	**0.3193** †	0.6483	**0.4753**	**0.3130**	**0.6522** †	**0.4761**	**0.3062**	**0.6170** †
cosine	0.4369	0.2740	0.5996	0.4369	0.2712	0.5630	0.4447	0.2877	0.5835	0.4496	0.2822	0.5363
dice	0.2965	0.1628	0.4668	0.3271	0.2002	0.4257	0.1956	0.0976	0.3469	0.2281	0.1350	0.3414
euclidean	0.2856	0.1680	0.3349	0.2626	0.1583	0.2946	0.3622	0.2147	0.4068	0.3325	0.1936	0.3620
baseline	0.4751	0.3018	0.6682	0.4691	0.2939	**0.6593**	0.4511	0.2893	0.5892	0.4551	0.2911	0.5691

test collection of initiatives, being stored as a bag of the described index terms.

To summarize an initiative, simple and bigrams terms are extracted from the initiative full text and are employed to query the Lucene index. The textual contents from the top scoring retrieved paragraphs are concatenated to build the initiative summary. The intuition behind this summarization method is that paragraphs most similar to the full text of an initiative are a good representation of the original text. In our experiments we have created automatic summaries drawing the 25%, 50% and 75% of the most representative paragraphs for each initiative.

With regard to subject assignment using our multilabel classification approach, in our experiments we have evaluated the use of Classifier Chains (CC) using two types of base classifiers. Runs labeled as CC-SVM employed SVM classifiers, using the Java version of LibSVM code with the default settings. The other set of CC runs, labeled as CC-NB, employed as base classifiers the Naive Bayes implementation provided by the Weka engine. In both cases we have exploited the facilities provided by the respective libraries to retrieve different kinds of scores which were used to measure the confidence of the predicted outputs.

The implementation of the two level hierarchical classification scheme described at the end of Section 3, which is labeled as 2LH-SVM in our runs, employs at each level a CC multilabel strategy using, again, SVM classifiers provided by LibSVM as elementary categorization models.

To assess the quality of the subject assignment provided by these classification schemes, we have employed the following label based performance measures (Tsoumakas et al., 2010): micro-averaged precision (MiP), micro-averaged recall (MiR) and micro-averaged F-value (MiF). In Table 1 we show the performance results obtained by the three multilabel categorization methods being employed in our experiment (CC-SVM, CC-NB and 2LH-SVM) using five alternatives to initiative text representation (full text, manually created summaries, automatic summaries using the 75%, 50% and 25% more relevant portions of text). In the reported results the number of predicted subjects retrieved from multilabel classifiers output was fixed to 10 after a preliminary test study, which showed that using this quite large value helps to capture the diversity of initiative contents.

The reported experimental results are apparently quite low, although we must consider that this is a relatively large multilabel classification problem, with 1405 labels actually occurring in our corpus (out of 5176 subjects in the EUROVOC thesaurus) and only 3607 training documents at each fold. The best results were obtained when full text initiatives were employed, making evident that our simple automatic summary scheme is not actually suitable to extract the most relevant passages, since the performance values decrease with the size of the summaries being used.

The performance values for the proposed multilabel categorization methods are fairly close, with CC-SVM being the best performing approach followed by 2LH-SVM and CC-NB. The lower performance shown by CC-NB may be due to the higher capabilities of SVM based classifiers, traditionally considered among the most competitive algorithms in text classification tasks. The difference between CC-SVM and 2LH-SVM is small, since in both cases the operation ultimately ends up being quite similar. We conjecture that lower performance is due to a typical issue with hierarchical classifiers using top-down strategies. These approaches are very sensitive to error propagation between local classifiers at different levels. False positives at first level tend to induce incorrect routes that worsen overall precision and false negatives block promising paths, affecting both recall and precision measures.

5.2 Profile Matching Evaluation

In the experiments assessing profile matching we have employed the CC-SVM multilabel scheme to create the profile of weighted subjects that will be matched against the MPs profiles. Initiative contents representation is accomplished using full text since this approach offered best results in the preliminary phase of experiments. The profile of each initiative is constructed by selecting the 10 subjects with better confidence scores predicted by the trained CC-SVM classifier. Each predicted subject in those constructed pro-

Table 3: Profile matching performance using SFIDF profiles.

	ideal			ideal with expansion			CC-SVM			CC-SVM with expansion		
	MAP	RPrec	R@10	MAP	RPrec	R@10	MAP	RPrec	R@10	MAP	RPrec	R@10
simple	**0.5282** †	**0.3467** †	**0.7036** †	**0.5029** †	**0.3231** †	0.6547	**0.4907** †	**0.3271** †	**0.6772**	**0.4777**	**0.3085** †	**0.6283** †
cosine	0.4391	0.2808	0.6057	0.4155	0.2662	0.5399	0.4434	0.2806	0.5909 †	0.4120	0.2628	0.5189
dice	0.2809	0.1556	0.4588	0.2897	0.1766	0.4000	0.2033	0.0994	0.3517	0.2254	0.1313	0.3342
euclidean	0.2185	0.1303	0.2152	0.2281	0.1356	0.2361	0.2563	0.1499	0.2570	0.2620	0.1605	0.2674
baseline	0.4751	0.3018	0.6682	0.4691	0.2939	**0.6593**	0.4511	0.2893	0.5892	0.4551	0.2911	0.5691

files was weighted by this confidence score.

We have used the following performance measures over the ordered list of candidate MPs created according to the scores provided by the matching functions described in Section 4.

MAP. Mean Average Precision is the mean average precision over the predicted MP assignments for the test set initiatives at each fold.

Average precision aggregates precision values at different points of the ranked list of predicted MPs where a relevant MP was predicted.

RPrec. R-Precision is the precision value computed at position R in the ranked list of candidates, being R the number of relevant MPs actually associated with the current initiative.

R@10. Recall value computed for the top 10 predicted candidates.

In order to have an estimate of the expected maximum performance of the proposed MP assignment scheme, we have evaluated the matching functions using a sort of "ideal initiative profile" built using the set of real EUROVOC subjects manually assigned by the parliament documentalists to the test initiatives. The performance values obtained by those "ideal initiative profiles" are labeled as *ideal* in Table 2 and Table 3.

Finally, as baseline we have employed a direct MPs assignment scheme using a multilabel classifier based on CC using SVMs as elementary classifiers (thus avoiding the use of MP profiles). The classifier that implements this direct assignment scheme uses as features the EUROVOC subjects associated to each test initiative, either the real (ideal) subjects coming from parliament documentalists or subjects predicted by the CC-SVM method. This direct classifier was trained to predict the set of relevant MPs using the real subjects of the initiatives in the training set.

Table 2 shows the results obtained using the SF profile generation method and Table 3 the results using SFIDF profiles, both methods were described in Section 2. With each MP profile generation strategy, the quality of the profile matching functions *simple*, *cosine*, *dice* and *euclidean* is assessed using manually assigned EUROVOC subjects to create the "ideal initiative profiles" and CC-SVM predicted subjects as initiative weighted profiles. In both cases, the contribution of similar subject expansion using Lin's taxonomy similarity is also checked. For each document profile generation strategy best results across the evaluated profile matching functions are shown in boldface. Statistically significant improvements (Teh, 2000) with respect to the employed baseline are marked with † in both tables, using Student's paired t-test with a significance level $\alpha = 0.05$.

The first direct conclusion drawn from those results is that the best matching function is clearly the simple measure (which simply adds the products of the weights of subjects shared by document and MP profiles). The second best matching function is the cosine similarity, whereas the euclidean distance and the weighted Dice coefficient get rather bad results. The fact that the other matching functions were unable to offer competitive results in comparison with the simple matching function leads us to the conclusion that normalization of profile weights or normalization of matching function values attenuate the actual discriminative power of the subjects that dominate the profiles being matched.

Regarding the profile construction methods, the best results depend on the matching function being considered: SFIDF is always better than SF when using the simple matching. However, SF tends to be better than SFIDF for both cosine and Dice, and it is always better in the case of the euclidean distance.

With respect to the use or not of the expansion with similar subjects, the differences are rather small, and the behavior is somewhat different depending on the performance measure: it is clearly preferable not to use the expansion from the point of view of the recall measure. For R-precision, it is also better not to use expansion if we consider the simple or the cosine matching functions, and the opposite is true for Dice and euclidean distance. This is also the tendency for the other performance measure, MAP.

The results of the baseline method are quite good, they are only surpassed by those of the simple matching function. Therefore, the configuration of our method that obtains the best results is to use the simple matching function together with the SFIDF profile construction method, and without using the expansion

with similar subjects.

If we compare the results obtained by the (normal) profiles with those of the "ideal profiles", we can observe that ideal profiles are better, as might have been expected. However, the differences are small. For example, for the best configuration, the differences between ideal and normal profiles are in percentage 7.1%, 5.6% and 3.7% for MAP, RPrec and R@10, respectively. Therefore, normal profiles perform close to the expected maximum performance offered by the ideal profiles. Thus, although the categorization performance of multilabel classifiers shown in Table 1 was not very high, using the subjects predicted by these classifiers during the MP profile matching have had an acceptable performance and it was consistent with the behavior of the ideal profiles based on actual thesaurus subjects assigned manually. This behavior along with the former effect of using the *simple* matching leads us to believe that the assignment of MP based on subjects profiles could be heavily dominated by the most frequent subjects, which are those which usually have higher weights in the SF and SFIDF schemes and get higher rates of success in the predictions made by our multilabel classifiers, because of the support of a greater range of positive examples characterizing them.

6 CONCLUSIONS

A content based filtering method to deal with the problem of assigning parliamentary documents to members of the parliament potentially interested on them has been described and evaluated. User and document profiles are defined using subjects taken from a conceptual thesaurus, and document profile generation is modeled as a multilabel categorization problem. The proposed method has been validated using real world data from a collection of parliamentary documents, manually annotated by human experts. Several matching approaches were evaluated and we were able to get an approximate document conceptual representation and a profile matching method achieving performance measures not very far from the "ideal" case. More work needs to be done in improving the applied multilabel categorization methods and also to evaluate alternative matching functions. Although *a priory* the similarity-based expansion of subject profiles seemed to be a promising alternative to get more flexible matching, the simple strategy we have proposed was unable to improve profile matching quality.

ACKNOWLEDGEMENTS

Paper supported by the Spanish "Ministerio de Economía y Competitividad" under projects TIN2013-42741-P and FFI2014-51978-C2-1.

REFERENCES

Belkin, N.J., and Croft, W.B. (1992). Information Filtering and Information Retrieval: Two Sides of the Same Coin? Communications of the ACM, 35:29–38.

de Campos, L.M., Fernández-Luna, J.M., Huete, J.F., Martin-Dancausa, C.J., Tur-Vigil, C., Tagua, A. (2009). An Integrated System for Managing the Andalusian Parliament's Digital Library. Program: Electronic Library and Information Systems, 43:121–139.

Chang, C.-C and Lin, C.-J (2011). LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27.

Gauch, S., Speretta, M., Chandramouli, A., and Micarelli, A. (2007). User Profiles for Personalized Information Access. In: The Adaptative Web. LCNS, vol. 4321, pages 54–89.

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

Hanani, U., Shapira, B., and Shoval, P. (2001). Information Filtering: Overview of Issues, Research and Systems. User Modeling and User-Adapted Interaction, 11:203–259.

Lantz, B. (2013). Machine Learning with R. Packt Publishing Ltd.

Lin., D. (1998). An Information-Theoretic Definition of Similarity. Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), pages 296–304.

Lops, P., de Gemmis, M., and Semerano, G. (2011). Content-based Recommender Systems: State of the Art and Trends. In: Recommender Systems Handbook, pages 73–105, Springer.

Pazzani, M., and Billsus, D. (2007). Content-based Recommendation Systems. In: The Adaptive Web. LCNS, vol. 4321, pages 325–341.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011). Classifier Chains for Multi-label Classification. Machine Learning, 85(3):333–359.

Silla Jr., C.N., and Freitas, A.A. (2011) A Survey of Hierarchical Classification across different Application Domains. Data Mining and Knowledge Discovery, 22(1-2):31–72.

Tsoumakas, G., Katakis, I., Vlahavas, I. (2010). Mining Multi-label Data. In Data Mining and Knowledge Discovery Handbook, pages 667–685, O. Maimon, L. Rokach (Eds.), Springer.

Yeh, A. (2000). More accurate tests for the statistical significance of result differences. In Proceedings of the 18th International Conference on Computational Linguistics (COLING), pages 947–953.