Comparing Machine Learning and Information Retrieval-Based Approaches for Filtering Documents in a Parliamentary Setting

Luis M. de Campos(✉), Juan M. Fernández-Luna, Juan F. Huete, and Luis Redondo-Expósito

Departamento de Ciencias de la Computación e Inteligencia Artificial, ETSI Informática y de Telecomunicación, CITIC-UGR, Universidad de Granada, 18071 Granada, Spain
{lci,jmfluna,jhg,luisre}@decsai.ugr.es

Abstract. We consider the problem of building a content-based recommender/filtering system in a parliamentary context which, given a new document to be recommended, can decide those Members of Parliament who should receive it. We propose and compare two different approaches to tackle this task, namely a machine learning-based method using automatic document classification and an information retrieval-based approach that matches documents and legislators' representations. The information necessary to build the system is automatically extracted from the transcriptions of the speeches of the members of parliament within the parliament debates. Our proposals are experimentally tested for the case of the regional Andalusian Parliament at Spain.

Keywords: Content-based recommender systems · Information filtering · Information retrieval · Machine learning · Parliamentary documents

1 Introduction

Politicians in general and Members of Parliament (MP) in particular, need to be concerned about the reality of the territory, region or country where they develop their activity. This is particularly true in relation to these matters more related with their specific political interests. For example, an MP who is specialized in educational issues or the health minister should be specially interested in receiving information concerning their respective fields of interest. However, at present, the amount of information that is generated and is available through the Information and Communication Technologies (ICT) is enormous, so it is not easy to decide what is interesting and what is not. As Shamin and Neuhold stated in [21], in the context of the European Parliament, "MPs need to be selective in their information input".

Let us consider a stream of documents that may be distributed among the MPs. These documents can be news releases, technical reports or parliamentary initiatives, for example. We would like to build an automated system able to

S. Moral et al. (Eds.): SUM 2017, LNAI 10564, pp. 64–77, 2017.
DOI: 10.1007/978-3-319-67582-4_5

recommend those MPs who should receive each document, taking into account both its own content and the specific interests and preferences of each MP.

Therefore, our research falls in the context of content-based recommender/filtering systems [10,18], which suggest items to users according to their preferences (represented by a profile or model of some kind), also taking into account some characteristics of the items (their textual content in our case). There are a lot of works addressing the recommendation/filtering problem in many domains and applications (see for example the three survey papers [3,16,17]). However, we are not aware of any such a system in a parliamentary context, except our own previous work [7,20]. Content-based recommender systems can be built using either information retrieval-based (IR) methods, which generate recommendations heuristically [1,2,9,15], or machine learning-based (ML) methods, mainly supervised classification algorithms for learning user models [4,5,12,13,19,22].

The objective of this paper is precisely to study and compare the capabilities of IR-based and ML-based methods in the parliamentary context we are considering. Therefore we propose two relatively simple approaches to create the recommender system, both based on first building a training document collection. One approach uses an Information Retrieval System (IRS) to explore this document collection, whereas the other uses this collection to generate a set of classifiers, one per MP. The training document collection will be obtained from the transcriptions of the speeches of the MPs in the parliamentary debates. The basic assumption is that these documents can provide information about the interests and preferences of the MPs. In order to compare our proposals, we shall perform experiments using a collection of MPs interventions from the regional Parliament of Andalusia at Spain.

The rest of the paper is organized in the following way: Sect. 2 gives details of the proposed IR-based and ML-based approaches to be compared. Section 3 contains the experimental part of the paper. Finally, Sect. 4 includes the concluding remarks and some proposals for future work.

2 Approaches for Recommending

The scenario that we consider is the following: we have a set of MPs $\mathcal{MP} = \{MP_1, \ldots, MP_n\}$. To the parliament documents arrive that must be distributed among the MPs according to their interests and preferences. We want to build a system that, given a new document, automatically selects those MPs that could be interested in reading it. Associated to each MP_i there is a set of documents $\mathcal{D}_i = \{d_{i1}, \ldots, d_{im_i}\}$, each d_{ij} representing the transcription of the speech of MP_i when participating in the discussion of a parliamentary initiative. The complete set of documents is $\mathcal{D} = \cup_{i=1}^{n} \mathcal{D}_i$. $\mathcal{D}$ is the training document collection that will be used by both the IR-based and the ML-based approaches.

2.1 The ML-Based Approach

The idea is simply to use the transcriptions of the speeches of the MPs in the parliamentary debates, $\mathcal{D}$, as training data to train a binary classifier (relevant/non-relevant) for each MP. Then, given a new document to be filtered/recommended, we use all these classifiers to decide which MPs should receive this document, namely those MPs whose corresponding classifier predicts the relevant class or, alternatively, assuming that the classifiers give a numerical output (a score) instead of a binary value, we could generate a ranking of MPs in decreasing order of score, thus recommending the document to those MPs whose score is greater than a given threshold.

In order to build a standard binary classifier for each MP we need training data (documents in this case), both positive (relevant documents) and negative (irrelevant documents). We shall consider that the own interventions/speeches of an MP are positive training data for building the classifier for this MP. Therefore, for each MP_i the set of positive examples is precisely $\mathcal{D}_i$. We shall also consider that all the interventions which are not from an MP are negative training data for the classifier associated to this MP. Hence the set of negative examples for each MP_i is $\mathcal{D} \setminus \mathcal{D}_i$.

2.2 The IR-Based Approach

In this case we are going to use the documents in $\mathcal{D}$ in two different ways to feed an Information Retrieval System (IRS). This IRS will be used to retrieve the documents that are more similar to the document to be filtered/recommended, which plays the role of a query to the system. The two ways in which $\mathcal{D}$ is transformed into an indexed document collection, which were originally proposed in [7], are the following:

The Collection of MP Interventions. The documents to be indexed by the IRS are just those in $\mathcal{D}$, i.e. all the interventions of all the MPs in the training set. In this case, what we obtain as the output for a query (which is the document to be filtered) is a ranking of documents, each one associated with an MP. Then we replace a document in the ranking by its associated MP. However, this new ranking of MPs may contain duplicate MPs with different scores (corresponding to different interventions of the same MP). In order to get a ranking of non duplicate MPs, we remove all the occurrences of an MP except the one having the maximum score. We call this approach IR-i.

The Collection of MP Profiles. To avoid the previous problem of having to remove duplicates from the ranked list retrieved by the IRS, another option is to group together all the interventions of each MP in only one document, thus obtaining a document collection with as many documents as MPs. More precisely, from each set $\mathcal{D}_i$ we build the single document $d_i = \cup_{j=1}^{m_i} d_{ij}$ and then use $\cup_{i=1}^{n} d_i$ as the document collection to be indexed by the IRS. In this case the

output of the system as response to a query is directly a ranked list of MPs. We call this approach IR-p.

In the two cases considered the system obtains a ranked list of MPs in decreasing order of score. Nevertheless, and due to efficiency considerations, an IR system does not compute the document-length normalization and as consequence the output scores vary with the number of terms in the query. Although these raw scores are valid for obtaining a MP's ranking (not to compute the length normalization does not affect the ranking, the final aim of an IR system) this is not the case for document recommendation purposes. Particularly, in this problem we are looking for a common threshold that should be used to recommend a document to those MPs whose score is greater than this value, independently of the query. In order to be able to determine such threshold, the raw scores are normalized by dividing by the maximum score. Note that in this case the normalized score represents a similarity percentage with respect to the top ranked MP.

3 Experimental Evaluation

The evaluation of our proposals will be carried out using all the 5,258 parliamentary initiatives discussed in the 8th term of office of the Andalusian Parliament at Spain[1], marked up in XML [8].

Each initiative contains, among other things, the transcriptions of all the speeches of the MPs who intervene in the debate, together with their names. There is a total of 12,633 different interventions, but we have only considered the interventions of those MPs who participate in at least 10 different initiatives, a total of 132 MPs. All the initiatives were preprocessed by removing stop words and performing stemming.

Regarding the evaluation methodology, we shall use the repeated holdout method [14]. Concretely, the set of initiatives is randomly partitioned into a training and a test set (containing in our case 80% and 20% of the initiatives, respectively), and the process is repeated (5 times in our case), thus averaging the results of the different rounds.

From the initiatives in the training set, we extract the interventions of all the MPs to form our training document collection $\mathcal{D}$. Then we build a classifier for each MP, following the ML-based approach (described in Sect. 2.1), and also an IRS (in the two ways described in Sect. 2.2) following the IR-based approach. In order to train a binary classifier for each MP_i from $\mathcal{D}_i$ and $\mathcal{D} \setminus \mathcal{D}_i$, we have used Support Vector Machines [6], which is considered as the state-of-the-art technique for document classification (we used the implementations of SVM available in R[2]). From the IR perspective, we have used the BM25 information retrieval model (using the implementation in the search engine library Lucene[3]), which is also a state-of-the-art technique in document retrieval [1].

[1] http://www.parlamentodeandalucia.es.
[2] https://cran.r-project.org.
[3] https://lucene.apache.org.

The initiatives in the test set are used as the documents to be filtered/recommended (using only the transcriptions of all the speeches within each initiative as the text of the document). We consider that each test initiative is relevant only for those MPs who participate in it. Notice that this is a very conservative assumption, since this initiative could also be relevant to other MPs interested in the same topics discussed in it, but it is the only way to establish a kind of "ground truth".

The evaluation measures used to assess the quality of the filtering/recommendation system are those typically used in text classification: we compute the precision, recall and the F-measure of the results associated to each MP_i. Precision is the ratio between the number of truly relevant test initiatives for MP_i which are correctly identified by the system (True Positives, TP_i) and the total number of test initiatives identified as relevant for MP_i ($TP_i + FP_i$, being FP_i the False Positives), $p_i = TP_i/(TP_i + FP_i)$. Recall is the ratio between TP_i and the number of test initiatives which are truly relevant for MP_i ($TP_i + FN_i$, being FN_i the False Negatives), $r_i = TP_i/(TP_i + FN_i)$ (see Table 1). Then we can compute the F-measure, as the harmonic mean of precision and recall, $F_i = 2p_i r_i/(p_i + r_i)$. To summarize all the measures, associated to each MP_i, we shall use both macro-averaged (M) and micro-averaged (m) measures [23]:

$$Mp = \frac{1}{n}\sum_{i=1}^{n} p_i \qquad Mr = \frac{1}{n}\sum_{i=1}^{n} r_i \qquad MF = \frac{1}{n}\sum_{i=1}^{n} F_i$$

$$mp = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FP_i)} \qquad mr = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FN_i)} \qquad mF = \frac{2mp\,mr}{mp + mr}$$

Table 1. Relations between TP_i, FP_i and FN_i with true relevance of the documents to be recommended and the scores.

	Truly relevant	Truly irrelevant
Score $\geq$ threshold	TP_i	FP_i
Score $<$ threshold	FN_i	TN_i

All the previous performance measures heavily depend on the selected threshold used to recommend the document to those MPs whose score is greater than this threshold. We will experiment with different thresholds, ranging from 0.1 to 0.9. It should be noticed that, as the scores obtained by the ML-based and the IR-based approaches represent different things (probability in one case and similarity with the best result in the other), the same happens with the thresholds.

We are going to also use another evaluation measure that does not depend on any threshold but it measures directly the ranking quality. This measure is the well-known in the IR field Normalized Discounted Cumulative Gain (NDCG) [11]. This evaluation metric tries to estimate the cumulative relevance

gain obtained by examining the first documents (MPs in our case) in a retrieved list of results. Since users tend to check only the first results, a discounting factor is used to reduce the document effect over the metric value as its position increases within the ranking. The metric value for a given list of MPs, is calculated as follows:

$$NDCG@k = \frac{1}{N} \sum_{i=1}^{k} \frac{2^{rel(d_i)} - 1}{\log(i+1)}, \tag{1}$$

where k is the number of results evaluated (10 in our experiments); i is the ranking position of the MP being evaluated; d_i is the MP at position i; $rel(d_i)$ is the relevance value of d_i (either 0 or 1 in our case); the normalization factor N is the DCG for the ideal ranking, where all the relevant results are located consecutively in the first positions of the ranking. With this normalization, the

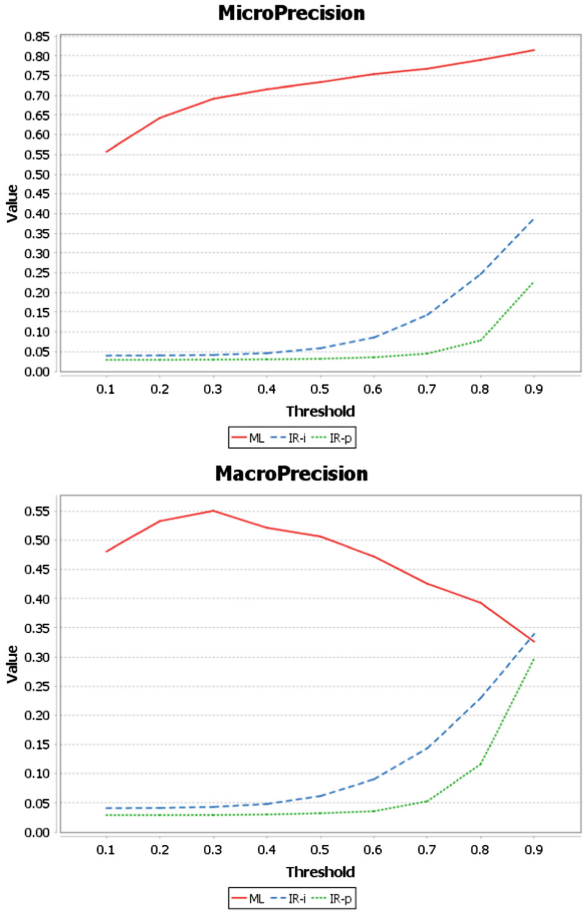

Fig. 1. Micro and Macro precision for ML, IR-i and IR-p using different thresholds.

metric values are always between 0 and 1, making it possible to calculate averages among different documents. This metric is computed for all the documents in the test set and then averaged.

3.1 Results

The results of our experiments for (macro and micro) precision, recall and F, using different thresholds (from 0.1 to 0.9) are displayed in Figs. 1, 2 and 3, respectively.

We can observe that, in general, the lower the threshold, the easier the system assigns the relevant value to documents, which increases the number of false positives and decreases precision. At the same time the number of false negatives decreases, thus increasing recall. When the threshold is high, the opposite

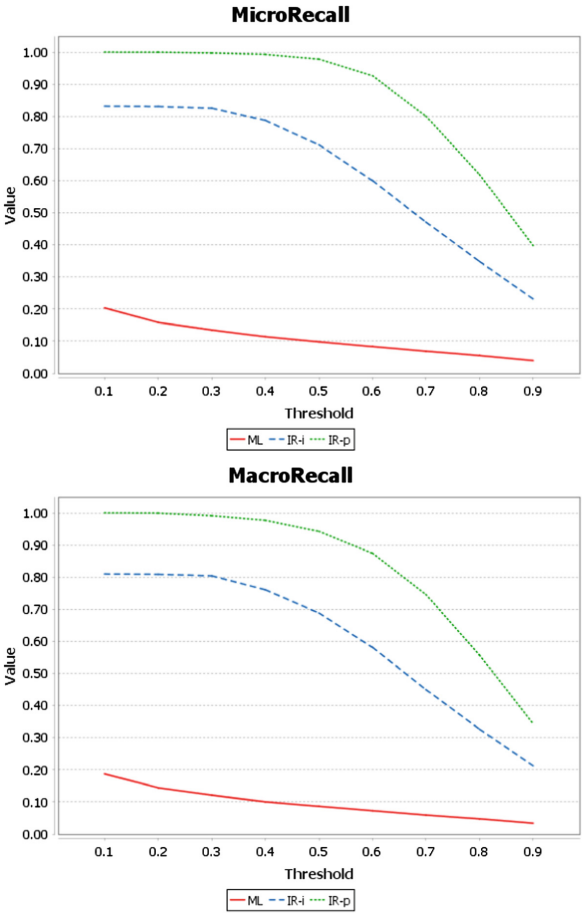

Fig. 2. Micro and Macro recall for ML, IR-i and IR-p using different thresholds.

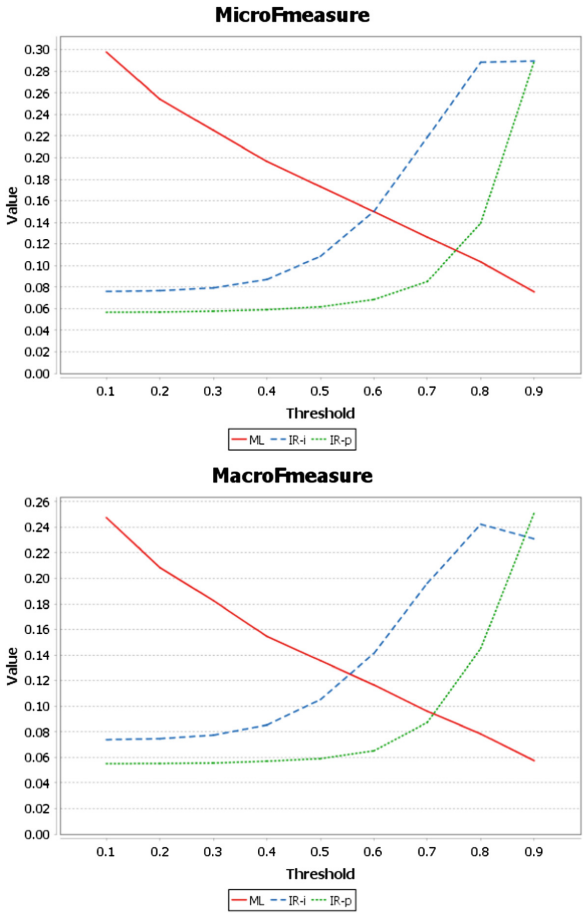

Fig. 3. Micro and Macro F measures for ML, IR-i and IR-p using different thresholds.

situation occurs, increasing precision and decreasing recall. The only anomaly to this general behaviour is with the ML-based approach and macro precision, which tends to decrease as the threshold increases. This may be due to a bad behaviour of this approach with those MPs having a low number of interventions (thus generating a poor training set), where the number of true positives decreases, even more steeply than the number of false positives, as the threshold increases (remember that with the macro measures all the MPs are equally important, independently on their number of interventions). More insights about this question will be given in the next section.

Nevertheless, the behaviour of the two approaches is quite different. The ML-based approach obtains relatively good precision values, much better than those of the IR-based approach. However, the recall values of the ML-based approach are very bad, whereas those of the IR-based approach are quite good.

Table 2. Best micro and macro F and NDCG@10 values obtained by ML, IR-i and IR-p.

Approach threshold	ML	IR-i	IR-p
	0.1	0.8	0.9
mF	0.2978	0.2896	0.2829
MF	0.2475	0.2423	0.2513
NDCG@10	0.6263	0.6246	0.6776

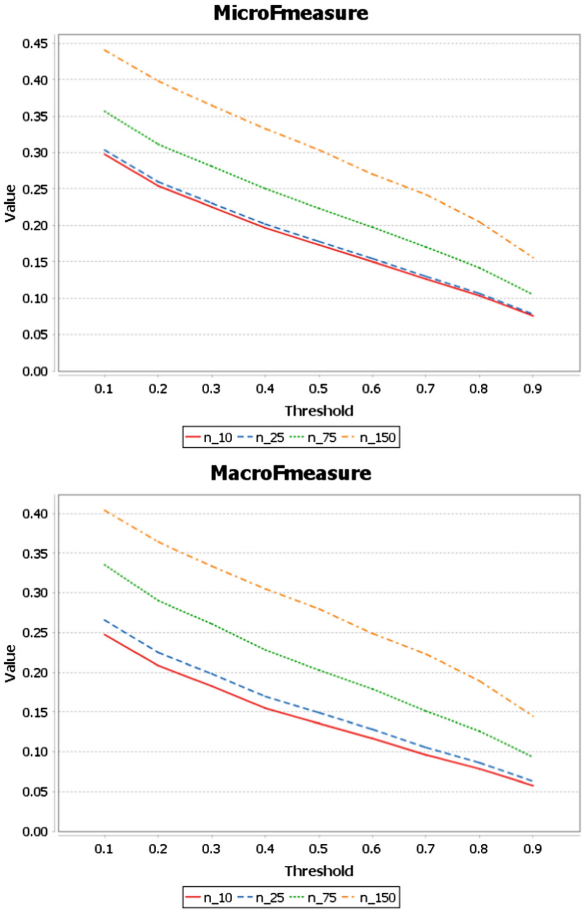

Fig. 4. Micro and Macro F measures for ML, using different thresholds and varying the minimum number of interventions.

The two IR-based approaches are quite similar, although IR-p gets more extreme values than IR-i (better in recall and worse in precision).

The F measure, which establishes a balance between precision and recall, clearly indicates that the ML-based approach works better with low thresholds and the opposite is true for the IR-based approach. However, there is no clear winner. Table 2 contains the best F values obtained by each approach, as well as the corresponding values of the NDCG@10 measure.

The values of mF and MF are very similar for the three methods, ML is slightly better in mF and IR-p is slightly better in MF. In fact a t-test (using the results of the five random partitions, and a confidence level of 99%) does not report any statistically significant differences between these methods. Concerning NDCG, a t-test indicates that IR-p is significantly better than both ML and IR-i, although there is no significant difference between ML and IR-i.

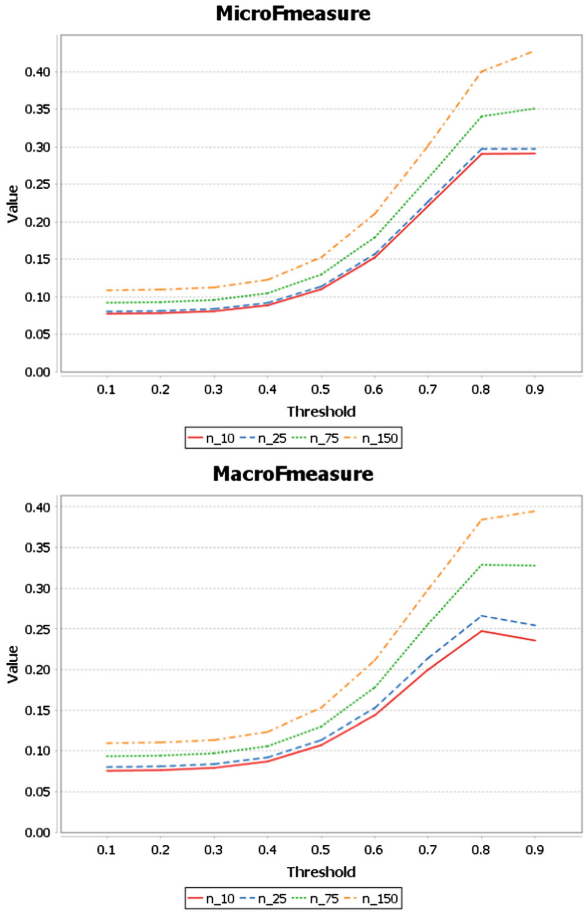

Fig. 5. Micro and Macro F measures for IR-i, using different thresholds and varying the minimum number of interventions.

3.2 Results When Varying the Number of Initiatives

As we said at the beginning of Sect. 3, the MPs being considered in this study are those who participate in at least 10 initiatives. This includes both MPs scarcely participating in the debates and other much more active (taking part in hundreds of initiatives). We want to evaluate the quality of the results depending on the number of initiatives where the MPs intervene.

To this end we have repeated our previous experiments, but fixing the minimum number of interventions of an MP which are necessary to include him in the study to greater values, concretely to 25, 75, and 150. Our goal is to evaluate whether a greater number of interventions of an MP translates into a better training set and hence to better results. For space reasons we do not include all the figures as we did in the previous experiments but only some of them for

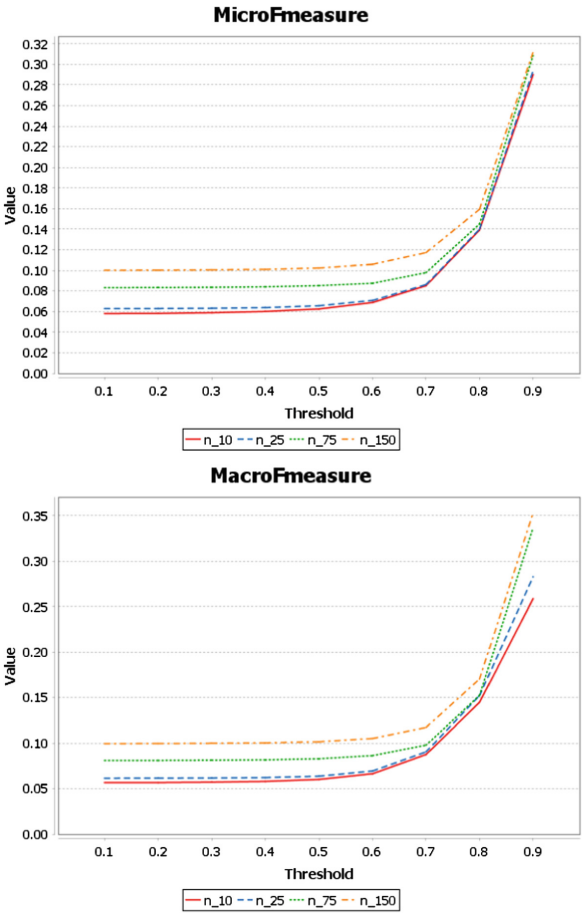

Fig. 6. Micro and Macro F measures for IR-p, using different thresholds and varying the minimum number of interventions.

illustrative purposes (micro and macro F for ML, IR-i and IR-p, see Figs. 4, 5 and 6, respectively).

As we can observe the trends are the same as in the previous experiments: in the ML-based approach the F measures decrease as the threshold increases, whereas the opposite is true for the IR-based approach. Moreover, the results are consistently better as the number of interventions required increases. Therefore, the two approaches could potentially reach better results if more training documents for each MP were available.

In Table 3 we show the best F values obtained for the different numbers of interventions, as well as the NDCG@10 values. For the F measures, the t-tests indicate that there are not significant differences between ML and IR-i in any case, whereas both ML and IR-i are significantly better than IR-p for micro F with sizes 75 and 150. For NDCG, again the differences between ML and IR-i are not significant but IR-p is significantly better than ML and IR-i with all the sizes.

Table 3. Best micro and macro F and NDCG@10 values obtained by ML, IR-i and IR-p, using different minimum numbers of interventions.

Approach	mF			MF			NDCG@10		
	ML	IR-i	IR-p	ML	IR-i	IR-p	ML	IR-i	IR-p
10	0.2978	0.2896	0.2829	0.2475	0.2423	0.2513	0.6263	0.6246	0.6776
25	0.3037	0.2971	0.2939	0.2658	0.2661	0.2829	0.6267	0.6242	0.6806
75	0.3568	0.3509	0.3085	0.3355	0.3288	0.3368	0.6132	0.6192	0.7086
150	0.4408	0.4282	0.3120	0.4039	0.3948	0.3532	0.5622	0.5744	0.6782

4 Concluding Remarks

In this paper we have proposed and compared two different approaches to build a system able to recommend/filter documents to the Members of Parliament. One approach is based on machine learning techniques, namely automatic document classification, whereas the other is based on information retrieval methods. The two approaches start from a collection of training documents composed of the interventions of the MPs in the parliamentary debates, which is assumed contains information about the interests and preferences of MPs. While the ML-based approach uses this collection to train a binary classifier for each MP, the IR-based approach uses an information retrieval system to index this collection and then retrieves the MPs which are more similar to the document to be recommended/filtered. In the two cases the output of the system is a ranked list (in decreasing order of score) of MPs. Then, given a fixed threshold, the system recommends the target document to those MPs whose score is above the threshold.

The two studied approaches behave quite differently in terms of recall and precision, and their best performance is attached using very different thresholds.

However, in terms of the F measures, both macro and micro (and to a lesser extent in terms of NDCG@10), the best results with both approaches are quite similar. Therefore, there is not a clear reason to prefer one approach to the other.

A possible weakness of the ML-based approach is that all the interventions which are not from an MP are considered as negative training data for the classifier associated to this MP. This is questionable: the interventions of other MPs which are about the same topics considered of interest for a given MP may be also relevant for him. For example an MP whose main area of interest is health could find interesting the interventions of other MPs also dealing with health. In this way the negative training data being used could contain positive data and this can limit the capacity of the classifier to discriminate between relevant and irrelevant documents. Therefore, we are interested for future research in using the so-called positive unlabeled learning techniques [24], which only assume the existence of a set of positive training data and a (usually larger) set of unlabeled data, but there is no negative training data.

Acknowledgements. This work has been funded by the Spanish "Ministerio de Economía y Competitividad" under projects TIN2013-42741-P and TIN2016-77902-C3-2-P, and the European Regional Development Fund (ERDF-FEDER).

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Boston (2011)
2. Belkin, N.J., Croft, W.B.: Information filtering and information retrieval: two sides of the same coin? Commun. ACM **35**, 29–38 (1992)
3. Bobadilla, J., Hernando, A., Fernando, O., Gutiérrez, A.: Recommender systems survey. Knowl. Based Syst. **46**, 109–132 (2013)
4. Billsus, D., Pazzani, M., Chen, J.: A learning agent for wireless news access. In: Proceedings of the International Conference on Intelligent User Interfaces, pp. 33–36 (2002)
5. Cohen, W.: Learning rules that classify e-mail. In: Papers from the AAAI Spring Symposium on Machine Learning in Information Access, pp. 18–25 (1996)
6. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, New York (2000)
7. de Campos, L.M., Fernández-Luna, J.M., Huete, J.F.: A lazy approach for filtering parliamentary documents. In: Kő, A., Francesconi, E. (eds.) EGOVIS 2015. LNCS, vol. 9265, pp. 364–378. Springer, Cham (2015). doi:10.1007/978-3-319-22389-6_26
8. de Campos, L.M., Fernández-Luna, J.M., Huete, J.F., Martin-Dancausa, C.J., Tur-Vigil, C., Tagua, A.: An integrated system for managing the andalusian parliament's digital library. Program Electron. Libr. Inf. Syst. **43**, 121–139 (2009)
9. Foltz, P., Dumais, S.: Personalized information delivery: an analysis of information filtering methods. Commun. ACM **35**, 51–60 (1992)
10. Hanani, U., Shapira, B., Shoval, P.: Information filtering: overview of issues, research and systems. User Model. User Adapt. Interact. **11**, 203–259 (2001)
11. Jarvelin, K., Kekalainen, J.: Cumulative gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. **20**, 422–446 (2002)

12. Kim, J., Lee, B., Shaw, M., Chang, H., Nelson, W.: Application of decision-tree induction techniques to personalized advertisements on internet storefronts. Int. J. Electron. Commerce **5**, 45–62 (2001)
13. Jennings, A., Higuchi, H.: A user model neural network for a personal news service. User Model. User Adapt. Interact. **3**, 1–25 (1993)
14. Lantz, B.: Machine Learning with R. Packt Publishing Ltd., Birmingham (2013)
15. Loeb, S.: Architecting personal delivery of multimedia information. Commun. ACM **35**, 39–48 (1992)
16. Lops, P., de Gemmis, M., Semeraro, G.: Content-based recommender systems: state of the art and trends. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P. (eds.) Recommender Systems Handbook. Springer, Boston (2011)
17. Lu, J., Wu, D., Mao, M., Wang, W., Zhang, G.: Recommender system application developments: a survey. Decis. Support Syst. **74**, 12–32 (2015)
18. Pazzani, M.J., Billsus, D.: Content-based recommendation systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web. LNCS, vol. 4321, pp. 325–341. Springer, Heidelberg (2007). doi:10.1007/978-3-540-72079-9_10
19. Pazzani, M., Billsus, D.: Learning and revising user profiles: the identification of interesting web sites. Mach. Learn. **27**, 313–331 (1997)
20. Ribadas, F.J., de Campos, L.M., Fernández-Luna, J.M., Huete, J.F.: Concept profiles for filtering parliamentary documents. In: Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, vol. 1, pp. 409–416 (2015)
21. Shamin, J., Neuhold, C.: 'Connecting Europe': the use of 'new' information and communication technologies within European parliament standing committees. J. Legislative Stud. **13**, 388–402 (2007)
22. Tjoa, A.M., Hofferer, M., Ehrentraut, G., Untersmeyer, P.: Applying evolutionary algorithms to the problem of information filtering. In: Proceedings of the 8th International Workshop on Database and Expert Systems Applications, pp. 450–458 (1997)
23. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook. Springer, Boston (2009)
24. Zhang, B., Zuo, W.: Learning from positive and unlabeled examples: a survey. In: International Symposiums on Information Processing, pp. 650–654 (2008)