Collaborative Document Classification: Definition, Application and Validation

Juan M. Fernández-Luna, Juan F. Huete, Guillermo Osorio
Dpto. Ciencias de la Computación e I.A. Universidad de Granada, 18071, Granada, Spain
{jmfluna,jhg}@decsai.ugr.es, gosorio@correo.ugr.es

***Abstract*—Document indexing is mostly a human task, where human indexers assign the most appropriate keywords to texts in order to represent or categorize their contents. It is usually performed as an individual manual task. In this paper we propose an extension where this process is enhanced with two main features: automatic classification, to support the knowledge of the expert, and collaboration between indexers, in order to obtain a more accurate result in the categorization. Then, we present a new approach called *Collaborative Document Classification*, describing their main elements and functionalities, as well as an application to the context of the political initiative indexing problem in the Andalusian Parliament. A computer simulation has been carried out with the aim of determining in a lab environment the possible benefits of this new approach, concluding that in several ways, the collaborative classification improves the indexing task.**

***Keywords*-collaboration; automatic classification**

I. Introduction

Document Classification or Categorization is the process by which a given document is labeled with one or more categories, which are representative of its content. These categories could be simple keywords or specialized descriptors from a predefined vocabulary (for example, a thesaurus). This process is also known as document indexing. Once a set of documents is categorized, the task of finding information is much easier as they are organized in such a way that similar documents belongs to the same categories.

This process can be performed either manually, by real users, experts or unskilled, or automatically, by means of classification algorithms [9]. Selecting one approach or another depends on several factors, as for example, the amount of information to be classified or the quality requirements in the resultant classification, among others.

This paper focuses on those situations where automatic categorization is configured as a computer-aided task for human experts, alleviating their work load, but at the same time when this process is critical in the sense that there is no room for error in the results. Therefore, it is necessary that an expert validates the quality of the automatic classifier's output, minimizing the risk of misclassified documents. For example, in official or critical documents generated by medical services or national parliaments.

But, another aspect that conforms the context for the research presented in this paper is the need of collaboration in the classification process. Let us suppose that an organization has created a library documentation service to index the documents that are generated. In such environments it is recognized that users usually work individually, applying some learned rules (maybe with the help of an automatic classification tool) that will adequately classify the majority of existing documents. There is a small number of documents, which is more difficult to classify, either because they are more ambiguous and therefore, harder to classify, or because they need more keywords to complete the classification (in multi-label classification problems). In such cases, it is then necessary that different individuals collaborate in order to find out the desired keywords. Then the problem is that there are no computer-supported tools that facilitate the collaboration among them and finally could help them to achieve the final decision.

The motivation for this research was born from the Andalusian Parliament, a regional chamber from Spain, where its librarians manually, and individually, index parliamentary initiatives by selecting a set of appropriate descriptors from the Eurovoc thesaurus. Our hypothesis is that supported by a computer application that facilitates collaboration among them, the indexing process would be more accurate.

Therefore, the objective of our research is to study document classification as a collaborative process: a set of individuals working together to select the most appropriate set of keywords representing the content of a document. The final contribution is a piece of software that implements relevant features borrowed from the Collaborative Information Retrieval (CIR) field [8], as sharing of knowledge and division of labour, applied to the document classification problem, validated by a simulation that reinforces the fact that collaboration is useful in this context. Specifically, division of labour is concerned with the task of assigning a set of jobs to a set of individuals, and knowledge sharing is used to allow communication between them and share information that other users can use to improve their classification.

The union of some parts of CIR and document classification originates a new research field that we have named *Collaborative Document Classification*, area that for the best of our knowledge is the first time that is presented in the specialized literature, configuring this proposal as the main contribution of this paper, which is organised as follows: In Section II, we describe in detail the problem that concerned us, explaining the motivation of this. Then, we explain how we pass from Collaborative Information Retrieval to Collaborative Classification in Section III, analysing the similarities and the differences with related work. The details of the Collaborative Classification Model presented in this paper are in Section IV and the evaluation of the system is presented in Section V. Finally, Section VI concludes the paper with implications and future works.

II. The Underlying Problem: Classification of Parliamentary Initiatives

As mentioned before, the motivation for this research came from the observation of the current process of indexing of political initiatives in the Andalusian Parliament by the staff of its Library Documentation Service.

Parliament works around the concept of parliamentary initiative, whereby an action taken by a member or political party is discussed in a plenary or specific area committee session. These initiatives are identified by means of an initiative code and are usually composed of a relative short textual description (the subject), plus a detailed body. They are also manually indexed with a set of labels that better represents its content. These labels must be obtained from a controlled vocabulary, more specifically descriptors from the Eurovoc thesaurus (a multilingual, multidisciplinary thesaurus covering the activities of the European Union).

Currently, each initiative contained in the incoming stream is assigned to any of the indexers, following no rules to produce this assignment. This means that all of them are able to index any initiative, regardless of the area to which they belong to, i.e. there are no specialized human indexers in agriculture, economics, education and so on, who could produce a more specific classification, taking the most of the possible expert knowledge.

Then, given an initiative, the human indexer, with a deep knowledge of the Eurovoc thesaurus, is able to assign one or more descriptors to it, which are the most appropriate according to its content (initially using the subject, and in case of any doubt, consulting the body for more information). But this is usually an individual process, in which few times the indexer asks for advice to other colleagues.

We may easily observe that there are three main problems in this process: (1) The indexing process is completely manual; (2) The staff is composed of "general" human indexers, without specialization in any field, and (3) Collaboration between indexers is almost null. Then, we think that this routine could be improved substantially, obtaining much better results and making it more efficient (1) being supported by an automatic classification tool that could help the human indexers by suggesting descriptors for each initiative that they could consider to index it; (2) having a specialized staff, where each indexer is expert in one area, so the indexing could be done with a finer granularity, and (3) collaborating more frequently with the rest of colleagues.

Another problem in the current work flow is the fact that the person who creates the initiative, does not worry about what descriptors, or more broadly speaking, keywords, she would use. This task is assigned to the human indexers. But a true fact is that the representation power of the author for expressing more accurately the content of the text is lost in some cases. Therefore, it can be highly convenient that this person, after writing the text of the initiative could also select the descriptors. Two are the benefits of this approach: on the one hand, the indexing process would gain in quality, as the initiative is indexed in its origin and, on the other hand, the workload of the indexers would be reduced considerably.

In order to allow indexing in origin, two problems must be consider: Firstly, we can not assume that the user has got any kind of knowledge about the vocabulary that she could use in this task, because she is not an expert on indexing and, secondly, the user does not know which are the rules of the organization to index (for instance, what is better the use of narrower or broader terms?). In order to solve these problems, the user could be helped by an automatic classification tool at the first moment, and collaboratively supported by the knowledge of the staff members, who know very well the indexing process. Therefore, there would be a second type of collaboration, in this case not only between the professionals, but also between them and non-experts.

III. Towards Collaborative Classification

Information Retrieval (IR), as defined in [1], refers to the representation, storage, organization and access of information. Traditionally, research in IR has focused on models of individual users but in the last years a new trend based on remotely teamworks, working together to satisfy a need for common information, and supported on advances in distributed technologies and computer hardware, is becoming stronger. Consequently, some researchers have realized that collaboration is an important feature which should be analysed in detail in order to be integrated with professional IR systems, upgrading these to Collaborative Information Retrieval (CIR) systems.

An early definition of CIR was given by S. Dumais et al. in [13] as *"any activity that collectively resolves an information problem taken by members of a work-team"*. P. Hansen and K. Järvelin [14] considered collaboration as an important component in the IR process, defining CIR as *"an information access activity related to specific problem solving activity that, implicitly or explicitly, involves human beings interacting with other human(s) directly and/or through texts (e.g., documents, notes, figures) as information sources in a work task related information search and retrieval process either in a specific workplace setting or in a more open community or environment"*.

CIR systems usually include some common features: session persistence, division of labour, knowledge sharing and awareness. *Division of labour* - Morris's survey in [12] describes ad hoc methods to avoid duplication of effort during a searching task, such as distributing the space of potential keywords, search engines or sub-tasks among different group members. *Sharing of knowledge* - In any collaborative setting, there will be a large and diverse knowledge base shared among groups of members. Each one will bring their own experience, expertise and topic knowledge to a particular searching task. What is needed is a way to enable the sharing of knowledge within the group [15]. *Group awareness* - Awareness is an essential element in distributed collaborative environments. Over the last decade, a number of researchers have explored the role of group awareness for supporting collaboration between distributed groups. [12]. *Session Persistence* - Storing a search session in a persistent format is a key requirement for facilitating collaboration during the session, revising the search at a later time, or sharing the results of a search with others [12].

Also CIR systems are divided into two types: synchronous and asynchronous. On the one hand, in the first class, teammates are able to interact between them at the same time; on the other hand, in the second type, the interaction is carried out in different time.

Particularly, in this paper, we have considered the *CIRLab* [6] framework. In general terms, it is a groupware framework, for experimenting with CIR techniques in different search scenarios. This framework has been designed applying design patterns and an object-oriented middleware platform to maximize its re-usability and adaptability in new contexts with a minimum of programming efforts.

The other main component of this research is Document Classification [9]. This area is also considered part of the IR and consists of assigning labels to a document, according usually to its textual content. There are three phases in the lifecycle of a text classification system, which traditionally have been addressed independently of each other: document indexing, text classifier learning and evaluation. The document indexing refers to the mapping of a document into a pattern that can be interpreted by the automatic classifier. In the second phase, the automatic text classifier learns from a set of categorized documents and learns the characteristics which define each class. Finally, the text classifier is evaluated to calculate its performance.

These two areas meet in a new one that we have called *Collaborative Document Classification*, in which we apply several features of CIR into the area of Document Classification. Specifically, techniques as division of labour or sharing of knowledge support a tool that helps human indexers to work collaboratively when classifying.

But this is not a mere and direct application from the CIR context. In this new Collaborative Document Classification area, two of the features previously mentioned take a different meaning. *Division of labour* - This is an important component in collaborative document classification, since in general terms it might have a great impact in the final process: The way in which the documents are distributed (divided) among the different individuals will have an effect in the background of the users, and as consequence, their usefulness in a collaborative framework. We analyse some algorithms because it is an important part of the collaborative classification. *Sharing of knowledge* - In our approach, the final decision about whether a keyword is appropriate or not will be lead by an individual, although in the process he/she can share his/her knowledge with others individuals, for instance by asking for keyword suggestion.

Therefore, and in practice, we have a division of labour phase before classifying and a sharing of knowledge phase after this process in order to improve the obtained classification. In Section IV, we show the details of the Collaborative Document Classification Model that we propose in this paper. Group awareness and session persistence take the same meaning as in CIR and are considered explicitly in the Collaborative Document Classification model.

There are few related works about collaborative classification. Most of them do not address the problem of text classification as a collaborative problem or do not use the term of collaboration with the meaning that we are using it, reason why we decide to focus on this topic. Collaborative classification is seen as social classification, i.e., the knowledge or the behaviour of users is used to classify something (documents, bookmarks, etc.). In this way, we mention [2] that proposes a methodology to carry out the collaborative classification idea of considering how similar users have classified a bookmark and [3] that compares a social classification using the uses of users with automatic extraction. This types of classification consist of extracting information about users and using it to classify. A similar work can be found in [11]. The authors propose an architecture that enables users to collaboratively build a faceted classification for a large, growing collection. But, the main difference with our work is that they considered collaboration as a set of individual user-machine collaborations, while we propose collaborations in the terms of user-user (direct communications between users instead of extracting the information from the global communications from the users to the system).

IV. Collaborative Document Classification

In this section, we are going to present the structure of the *Collaborative Document Classification Model* that we propose in this paper, which is graphically represented in Figure 1. The input is a continuous stream of initiatives that have to be indexed and the output is a set of labels (descriptors of a thesaurus) that are assigned to each initiative. Each one will be dispatched to a given human indexer (division of labour), who is responsible for its final classification. Then, the human indexer (using her knowledge, the support of an automatic classifier and the expertise of her colleagues (sharing of knowledge)) will compile the set of the most appropriate labels for representing the content of the initiative. Let us view in detail the different components:

Dispatcher Module: When the initiatives arrive to the system, they have to be distributed among the different

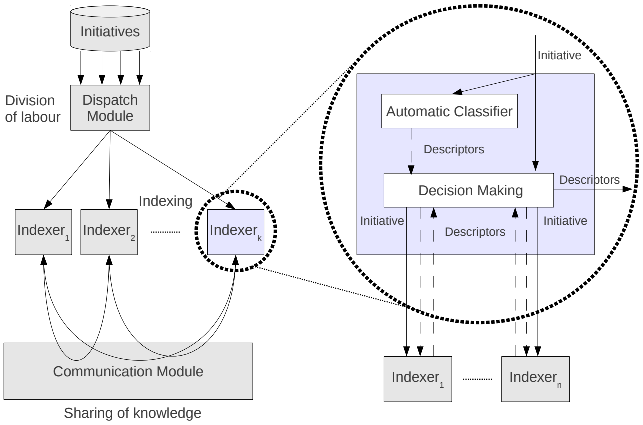
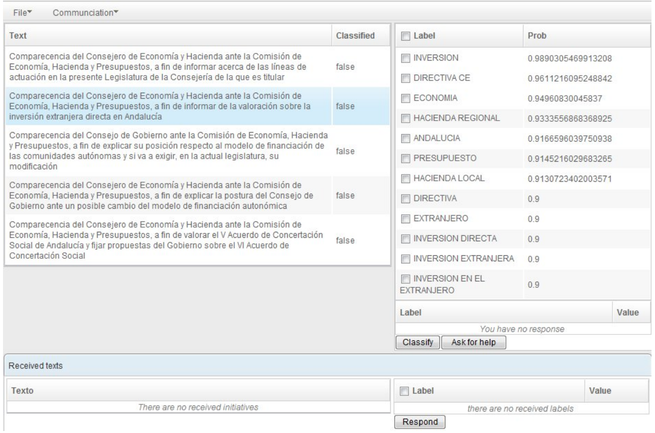


Figure 1. Collaborative Classification Model



Text	Classified
Comparecencia del Consejero de Economía y Hacienda ante la Comisión de Economía, Hacienda y Presupuestos, a fin de informar acerca de las líneas de actuación en la presente Legislatura de la Consejería de la que es titular	false
Comparecencia del Consejero de Economía y Hacienda ante la Comisión de Economía, Hacienda y Presupuestos, a fin de informar de la valoración sobre la inversión extranjera directa en Andalucía	false
Comparecencia del Consejo de Gobierno ante la Comisión de Economía, Hacienda y Presupuestos, a fin de explicar su posición respecto al modelo de financiación de las comunidades autónomas y si va a exigir, en la actual legislatura, su modificación	false
Comparecencia del Consejero de Economía y Hacienda ante la Comisión de Economía, Hacienda y Presupuestos, a fin de explicar la postura del Consejo de Gobierno ante un posible cambio del modelo de financiación autonómica	false
Comparecencia del Consejero de Economía y Hacienda ante la Comisión de Economía, Hacienda y Presupuestos, a fin de valorar el V Acuerdo de Concertación Social de Andalucía y fijar propuestas del Gobierno sobre el VI Acuerdo de Concertación Social	false

□ Label	Prob
□ INVERSION	0.9890305469913208
□ DIRECTIVA CE	0.9611216095248842
□ ECONOMIA	0.94960830045837
□ HACIENDA REGIONAL	0.9333556868368925
□ ANDALUCIA	0.9166596039750938
□ PRESUPUESTO	0.9145216029683265
□ HACIENDA LOCAL	0.9130723402003571
□ DIRECTIVA	0.9
□ EXTRANJERO	0.9
□ INVERSION DIRECTA	0.9
□ INVERSION EXTRANJERA	0.9
□ INVERSION EN EL EXTRANJERO	0.9

Label	Value
You have no response	

Texto
There are no received initiatives

□ Label	Value
there are no received labels	

Figure 2. Sharing of Knowledge

human indexers as they will be in charge of the categorization task. This is related to division of labour, a key component of the Collaborative Document Classification process: The way in which the work is distributed among the indexers will determine the background knowledge of the users, the type of communications and also the quality of the classification itself, as we will demonstrate in Section V. Different scenarios can be considered, going from specialized distribution, where an individual only indexes those initiatives in a given field (so we assume that she is expert on that field), to random distribution, where all the indexers have the same probability to index a given initiative (so we assume that the indexers have got a general knowledge about all the possible fields). Nevertheless, it might be considered other factors besides human expertise as, for example, the indexers' workload balance.

Indexing Module: Once the indexer receives an initiative, she starts the classification task itself. The final result is an ordered list of descriptors, sorted by the degree of aboutness of each keyword with respect to the initiative, represented by means of a weight (to what extent the descriptor is suitable for representing the content of the initiative).

Although the indexer can work with no help, in our model we have an additional module for automatic classification that supports the indexer's work by recommending a set of weighted labels for each initiative. This is particularly necessary for those users which have not been trained as indexers (the keywords are proposed in origin).

The automatic classifiers have to be trained in order to learn the classification models necessary to recommend further labels. For this purpose we can use a set of pre-classified initiatives, but also after a new initiative is finally classified, it is also used as training input for the automatic classifier associated to its responsible indexer. This helps to fit the automatic classifier to the particular background of its associated indexer. As mentioned before, this process is influenced by the division of labour strategy (dispatch module) since automatic classifiers behave different depending on the training data, which finally depends on the used division strategy. For example, the classification model learned from a set of heterogeneous initiatives (very different topics and randomly assigned to the indexer) is different from that one trained with homogeneous initiatives (all of them framed in a specific field where the indexer is specialized in).

Communication Module: The last phase in the process is referred as sharing of knowledge, part of CIR that allows indexers to communicate between them to work collaboratively during the classification. Sometimes, an indexer has problems to assign all the relevant labels to an initiative, probably because she is not an expert on a topic related to the initiative. In these cases, the indexer has to ask for help to the others colleagues to obtain extra information that helps her to make decisions about the relevant labels, as human knowledge and experience is very valuable in this task. Even though they are supported by automatic classification tools. This is the utility of the communication (sharing of knowledge) module (Figure 2 shows the interface that represents a chat box for synchronous communications).

In this sense, the source indexer sends, synchronously or asynchronously, to her workmates the initiative and the set of selected keywords. Then, they classify the initiative, also with support of their automatic classifiers, and propose a new set of labels to the source indexer who evaluates them. Thereby, an indexer could improve his classification with extra information. Obviously, in complex initiatives, all the indexers can collaborate by means of the sharing of knowledge component in order to get the final classification. This process can be iterated until the initiative has been classified with a high degree of satisfaction. Then, it is reported to all the indexers allowing a global vision of the progress in the classification task. There is no protocol for solving conflicts as the indexer in charge of the initiative is who makes the final decision.

ISBN: 978-1-61208-206-6

A. Implementation Details

We have shown the details of the Collaborative Document Classification Model designed in order to improve the performance of the indexers of the Andalusian Parliament. In this section we shall present some implementation details.

Dispatching strategies: The particular strategy used to distribute the initiatives among the different indexers will depend on the typology of the working environment. We have considered the following two different frameworks: (1) *Generalized indexing framework.* In this case we are assuming that all the indexers are equivalent for indexing purposes. Then, the context of the initiative is not relevant to determine the indexer in charge of its classification. Therefore, we can use a *Round-Robin* algorithm to distribute the initiatives sequentially between the different indexers. Note that this is the usual strategy used in many organizations, and particularly, this is the one used in the Andalusian Parliament. (2) *Expertise indexing framework.* In this environment, each indexer can be considered as an expert in one particular area (for instance, health services, economy, agriculture, etc.). Therefore, it seems natural that she is the responsible for those initiatives under her field. In order to distribute the initiatives in this environment it might be necessary that a human indexer read all the initiatives, choosing among the indexers the most appropriate candidate for assigning the final descriptors. Then, dispatching will become a critical process, so in this paper we propose a different alternative: This task can be done automatically using the content of the initiatives to select one of the high-level indexing areas.

For this last, purpose we propose the use of the K Nearest Neighbours (KNN) classification algorithm [9], which might be trained using a set of initiatives belonging to the different categories. Note that although some dispatching (classification) error can happened, it can be easily mitigated by the indexer (whenever she is not able to classify properly an initiative, she can use the share of knowledge module to re-distribute it to the proper indexer). Particularly, the KNN algorithm calculates the distances between the initiative to classify and the initiatives in the training set. Then, these initiatives are sorted by ascending order and the k nearest initiatives are selected. The algorithm classifies the new initiative in the category that most appears in the set of the k nearest initiatives (euclidean distance) or randomly if tie.

Helping the human indexers: An automatic classification tool: Independently of the working indexing environment, whenever the indexer receives an initiative she proceeds finding out the proper descriptors from a controlled vocabulary. This process can be done with the help of an automatic classification tools, which proposes a set of candidate descriptors. As mentioned before, this tool is particularly beneficial when the descriptors must be selected in origin. In our approach, this component is implemented using the *REBAYCT* algorithm [4][5] that uses Bayesian Networks for hierarchical text classification in a supervised and non supervised way. (Detailed information about how this hierarchical text classifier works could be seen in the already cited reference). The list suggested by the automatic classification module is evaluated by the human to determine which are relevant for the target initiative, so she could select the most appropriate and add those which she judged also relevant from her point of view. Each indexer has got her own automatic classifier which receives, in a feedback process, those initiatives which have been previously classified by the human as training data, so the automatic classifier can be adapted to the indexers preferences and learn new rules to find out the proper descriptors. Moreover, the automatic classifier is fed with the last initiatives already classified by the human, after the initial training phase, so it is up-to-date and totally adapted to the indexer.

Sharing of knowledge: With respect to the implementation, we have to mention that for the interactive communications between users, we have borrowed the middleware-based architecture used in *CIRLab* [6] framework as it is more appropriate for collaborative applications than a client/server architecture. This software was designed to develop CIR applications. The software implements all the CIR features, although for this application we only use the part related to the sharing of knowledge. The framework provides us many communication techniques such as sending synchronous messages, notifications of user connections, etc. In addition, we can create different collaborative working sessions so users can work in independent groups. In our case, we have integrated instant messaging between users, supporting the sharing of knowledge that concerns us in this application, i.e., the sharing of initiatives and labels. incorporates *CIRLab*.

V. Evaluation

We have developed a working prototype including all the already mentioned capabilities. The indexer can use the prototype to `search` for the best set of descriptors for a given initiative. Thus, in case of doubts, she can ask for help to the others colleagues (sending a package containing the initiative and a set of descriptors). Each colleague evaluates the proposal and, in case of having some additional labels which might be used to describe the initiative, decides to send them back to the original indexer who has to evaluate this new set of labels to obtain the final descriptors.

Nevertheless, changing the workflow of a (large) organization is difficult, and moving away from isolated to collaborative classification represents an important challenge that has to be evaluated properly. In this sense, we have designed a simulation study with the aim of demonstrating to the organization that working collaboratively and coordinately can improve the overall process.

A. Experimental Design

In our study, we use the same workflow, but replacing the indexer's search of the best set of descriptors by a process in which the indexers only judge those descriptors proposed by their associated automatic classifiers, as described in Section IV. By means of this simplification we can simulate different indexers working in an isolated and collaborative environments: (1) *Indexers working isolately*: Each indexer, i_i, judges an initial set of descriptors, i.e. the top-k descriptors proposed by her associated automatic classifier. (2) *Indexers working collaboratively*: Each indexer, i_i, judges an initial set of k descriptors proposed by her associated automatic classifier. Then, she sends a package to all the indexers, $i_j \neq i_i$. (A different alternative might be to select some of them, but we are focusing on measure the effect of working in a fully collaborative framework.) Each collaborative indexer, i_j, uses her own automatic classifier to obtain also a set of k descriptors. This indexer does not judge whether they are relevant or not for the initiative, but in case of having new descriptors (descriptors(i_j)\descriptors(i_i) $\neq \emptyset$), she sends a return package that will be evaluated by i_i to obtain the final descriptors.

B. Data Sets

In this experimentation, we have considered the initiatives discussed by different committee sessions in the Andalusian Parliament. These committees are usually attended by a reduced number of Members of Parliament (MPs) according to different areas of interest (agriculture, economy, education, etc.). Each initiative contains a subject, a text describing the content, plus a development, i.e. the full transcriptions of all the speeches discussing the concerned topic, although we are only taking into account the text from the subject.

We have two different set of initiatives: The first one, with 317 initiatives, where for each initiative we also know the particular committee session in which it has been discussed. This set is only used to learn how to distribute an initiative by the division of labour module (when necessary).

The second set has 7933 initiatives and is used for evaluation purposes. In this dataset, each initiative contains a subject and the set of descriptors from the Eurovoc thesaurus already assigned by the indexers, important information for validating our approach. This set will be split into training (80%) and test (20%) (none of the initiatives used for training will be then used for test). Training initiatives will be used as inputs for the different *REBAYCT* classifiers.

C. Indexers typologies and working settings

In our study, we shall consider three different indexers typologies which can work in three different working scenarios: the first type of user (U_1) represents a non-specialized indexer, which has a varied outlook of the parliamentary domain; the second type (U_2) is an indexer specialized in narrow or restrictive domains and the last one (U_3) represents an specialist who is in charge of the initiatives in a broader domain. Then we represent to the professional indexers working in the Parliament, as well as those users who could generate initiative. In order to get the area of expertise specialized for the human indexers we have considered the topics of the 6 main committee sessions in the Andalusian Parliament, i.e. the commissions of Economy, Agriculture, Education, Employment, Culture and Public Administration and Justice. The size of committee (in number of initiatives) will be useful to represent the different specialists. Thus, in the case of U_1, each individual *REBAYCT* classifier is trained with 1325 random initiatives; for U_2, the automatic classifiers are trained with 265 initiatives, while for U_3, the training set is composed of 1705 initiatives.

Related to the simulation of the working environments we will consider as baseline the situation in which the indexers are working isolated ($C1$) but also two collaborative settings: the first one, where an indexer works collaboratively in a non-specialized environment, i.e. the rest of the indexers are non-specialized ($C2$), and the second one, where the indexer works with specialists ($C3$).

D. Evaluation metrics

Our aim is to determine the effect of using a collaborative approach in different indexing scenarios. In order to evaluate our approach, we are going to consider two different criterion: On the one hand, the quality of the final classification and, on the second hand, the utility of the communications.

The measurement of classification system's performance relies on two metrics very well known measures in the field of text classification: precision and recall. Precision gives us an estimate of how many of the found descriptors are relevant to the initiative whereas recall estimates how many of all the descriptors relevant for a given initiative were successfully found. The first one is not relevant in our experimentation: We assume that all the descriptors finally selected by the indexers must be relevant for the initiative. The second one, recall, is particularly interesting for our proposal since, in some way, it gives an idea of the number of relevant descriptors which might be found thanks to the collaborative support. Thus, we calculate the recall values for the three type of users measured as the proportion of the relevant descriptors among those suggested to the indexer (by her associated automatic classifier or the rest of her collaborative colleagues) relevant for the initiative:

$$recall = \frac{N^o \textit{proposed relevant descriptors}}{N^o \textit{total relevant descriptors}}. \qquad (1)$$

The second criteria to evaluate refers to the utility of communications (CU). So, in order to measure the utility of the communications we propose to consider the proportion of return packages (those including new labels, and therefore being able to increase the knowledge of the indexer) with respect to those sent. Note that in this measure we are

ISBN: 978-1-61208-206-6

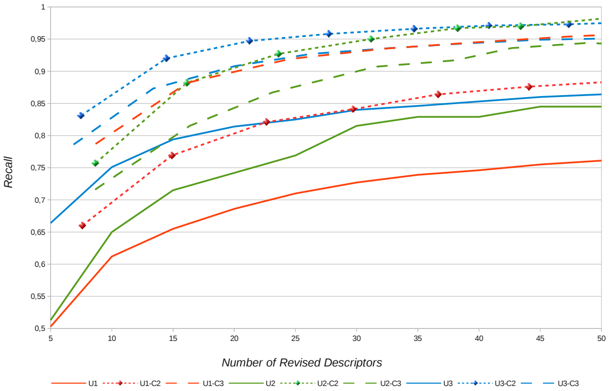


Figure 3. Recall values for the different environments

not considering whether the proposed descriptors are finally selected by the original indexer or not. This is a metric that we have specifically designed considering this problem.

E. Experimental Results

In this experimentation we consider the utility of the user from two different points of view: On the one hand, we shall measure how useful the communications that an indexer receives (incoming packages) are and, on the other hand, we shall determine the usefulness of the communications that an indexer sends to their colleagues (outgoing packages).

C_2	U_1		U_2		U_3	
k	C2	C3	C2	C3	C2	C3
5	10.339	14.679	14.679	14.528	9.906	7.489
10	9.796	12.332	12.264	12.716	8.909	6.739
15	10.188	12.686	11.509	10.83	8.316	6.739
20	9.713	12.196	11.169	11.547	7.724	6.299
25	9.328	11.818	10.603	10.452	7.753	6.228
30	9.381	11.622	8.943	8.452	7.208	5.753
35	9.147	11.486	9.32	7.886	7.043	5.689
40	9.313	11.026	9.396	7.66	6.785	5.495
45	9.049	10.913	8.603	7.396	6.668	5.624
50	9.192	10.815	8.792	7.471	6.504	5.36

Table I
USEFUL COMMUNICATIONS RECEIVED BY AN INDEXER

1) Utility of incoming communications: Table I shows, for each user, the usefulness of the communications under different scenarios. Some conclusions can be obtained: The first one is that the working environment (note that this is finally related to the division of labour strategies) has an effect on the utility of the communications. Thus, a non-specialized human indexer, $U1$, obtains more feedback when working in a specialized environment, whereas an indexer specialized in broad domains, $U3$, obtains more feedback when working in a general scenario. This can be explained because in $C3$ the indexers are specialists in their respective domains, so they are not useful for a specialist but they might be helpful for a non-specialized indexer.

The situation is quite different for the indexer specialized in a narrow domain. In this case, there is no difference if she works in a general or specialized environment. The large amount of useful communications, particularly if we compare with the other specialist $U3$, can be explained because her associated automatic classifier has been trained with less data, being not able to find some general rules learned by the others models. These rules suggests descriptors that could be also applied in this field (independently if these descriptors are valid for the initiative in a narrower domain).

Related to this last point is the study of whether the communications are fruitful or not. In this sense, a fruitful communication will help to find more proper descriptors, and as consequence it will improve the recall metric. Figure 3 can help to understand this situation. Particularly, we show the recall curves obtained under the different scenarios when considering the total number of descriptors analysed by an indexer. We use with solid lines to represent the results obtained when the user works isolated, the lines with dots and diamonds represents the recall values obtained when working in a general collaborative environment $C2$ and, finally, the dashed lines represent the results obtained in a specialized collaborative scenario, $C3$. The first conclusion is that in collaborative scenarios the recall increases considerably, which means that the communications are helpful in all the situations (it is better for the user to ask for help than to keep exploring the descriptors isolated).

Moreover, specialized indexers working isolated obtain best results than non-specialized ones (for $U1$ it might be applied the idea of "jack of all trades, master of none"). Now, let us focus on the situation presented when the indexers work in a collaborative environment. In this case, the specialized ones ($U2$ and $U3$) work much better when working with generalist indexers. The explanation of this fact is found in Table I, because non-specialized indexers propose more useful descriptors. This is particularly true for those specialists in narrow domains, as $U2$. But the situation changes for non-specialized indexers, being preferable to work in specialized scenarios. The reason seems to be clear: asking for help to similar people is good, but we will obtain much reliable results if we take into account the opinion of the specialist. In this sense, we would like to highlight that a general indexer working with specialists can obtain the same recall as the specialists. This results is particularly relevant because it supports the idea that in a collaborative environment it is possible to index at the origin (by the people who proposed the initiative) without worsen the quality of the indexing processes.

Finally, a last conclusion is that it will be convenient that non-specialized and specialized indexers work together in collaborative environments. This result is important because it opens a new research opportunity, i.e. we have to study carefully the way in which the division of labour is performed in the initial steps and its effects in the final results.

2) Utility of outgoing communications: In this section we shall discuss how useful is this particular indexer for the rest of her colleagues, analysing how many useful communications proposes. In Table II we show the obtained results. From this table we can see that, independently of the indexer typology, a greater number of communications is obtained when working in a non-specialized scenario, being the difference particularly relevant when assuming that the original indexer is a specialist. This corroborates the fact that specialized users are helpful in general environments and that broad-domain specialists collaborate more actively. With respect to non-specialized users, it is particularly relevant the large volume of useful communications in a specialized scenarios, becoming more valuable users (they can provide some kind of information that the specialized indexers, focused on a specific domain, are not able to capture).

C_2	U_1		U_2		U_3	
k	C2	C3	C2	C3	C2	C3
5	19.841	18.035	14.679	10.384	23.841	15.974
10	19.328	15.066	12.271	7.99	20.92	13.638
15	18.618	14.659	10.407	6,757	19.645	12.929
20	18.581	13.445	10.211	6,784	19.124	12.084
25	18.581	12.46	9.894	6,001	18.581	11.428
30	17.652	12.263	9.335	5,414	18.316	11.221
35	17.071	11.284	9.252	5,57	17.577	10.943
40	16.83	11.201	8.898	5,433	16.883	9.82
45	16.505	11.271	8.505	5,499	17.215	9.73
50	16.883	10.654	8.03	5,544	17.283	10.008

Table II
UTILITY OF USERS IN A COLLABORATIVE ENVIRONMENT

VI. CONCLUSION AND FURTHER RESEARCH

This paper has presented a new approach for document classification, based on collaborative techniques borrowed from CIR. The motivation for this development was the improvement of the current manual indexing of political initiatives from the Andalusian Parliament. This new technique is based on two important CIR concepts: division of labour and sharing knowledge. Beside, automatic classification is included to support the human indexing process. Then the individual and manual indexing is transformed to a collaborative task, where all the human indexers could be involved to improve the keyword assignation.

In order to determine if this approach could work in a real environment, we have designed and performed a simulation in the context of indexing political documents in this regional chamber, where several scenarios have been represented, considering the expertise of the human indexers (general knowledge about the source; specialized in a broad domain and in a narrow domain) and the working environment where they have been included (working isolated, with general knowledge indexers or with expert ones).

The main conclusion drawn from this evaluation is that collaboration substantially improves the classification process for all types of users in any working setting. More specifically, a non-specialized human indexer would obtain better results when they work with specialized colleagues (non-expert user working collaborative with expert users is able to classify as well as they are), and the other way around: a specialized indexer would improve her performance working with a set of non-specialized workmates.

With respect to further research, the first step is to evaluate the collaborative classification tool with real users by means of a user study in the Andalusian Parliament, once we have shown that the simulation shows these very interesting and good results. Also we are planning to test our model in other contexts, with different problems, for example, indexing of medical articles, where the number of documents is much higher as well as the specialization degree of the indexers.

Acknowledgement. This work was jointly supported by the Spanish Ministerio de Educación y Ciencia (TIN2008-06566-C04-01 and TIN2011-28538-C02-02), Spanish Research Program Consolider Ingenio 2010: MIPRCV (CSD2007-00018) and Junta de Andalucía Excellence Project TIC-04526.

REFERENCES

[1] Baeza-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] Benz, D. and Tso, K. and Schmidt-Thieme, L. *Automatic Bookmark Classification: A Collaborative Approach*. 2nd Workshop on Innovations in Web Infrastructure, 2006.

[3] Bogers, T. and Thoonen, W. and van den Bosch, A. *Expertise classification: Collaborative classification vs. automatic extraction*. SIG Classification Research Workshop, 2006.

[4] de Campos, L. M., Fernández-Luna J. M., Huete J. F., and Romero A. E., *Automatic Indexing from a Thesaurus Using Bayesian Networks: Application to the Classification of Parliamentary Initiatives*. LNCS 4724, pp. 865-877, 2007.

[5] de Campos, L. M. and Romero A. E., *Bayesian Network Models for Hierarchical Text Classification from a Thesaurus*, IJAR, 50(7), pp. 932-944, 2009.

[6] Fernández-Luna J. M., Huete J. F., Pérez Vázquez, R. , and Rodríguez-Cano J. C. *CIRLab: A groupware framework for collaborative information retrieval research*. IP&M, 44(1), pp. 256-273, 2009.

[7] Joho, H., Hannah, D., and Jose, J.M. *Revisiting IR Techniques for Collaborative Search Strategies*. ECIR, pp. 66-77, 2009.

[8] Morris, M.R. and Teevan, J. *Collaborative Web Search: Who, What, Where, When, and Why?* Morgan & Claypool, 2010.

[9] Sebastiani, Fabrizio. *Machine learning in automated text categorization*. ACM Comp. Sur., 34(1), pp. 1-47, 2002.

[10] Shah, C., and González, R. *Evaluating the Synergic Effect of Collaboration in Inf. Seeking*. SIGIR, pp. 24-28, 2011.

[11] Wu, Harris and Zubair, Mohammad and Maly, Kurt. *Collaborative classification of growing collections with evolving facets*. Hypertext, pp. 167-170, 2007.

[12] Morris, M. R. *A survey of collaborative web search practices*. CHI, 2008, pp. 1657–1660. 2008.

[13] Dumais, S., Grudin, J., Bruce, H., Fidel, R., Poltrock, S., and Pejtersen, A. M.,*Collaborative information retrieval (cir)*.The New Review of Inf. Behaviour Res., pp. 235247, 2000.

[14] Hansen, P. and Järvelin, K. *Collaborative information retrieval in an formation-intensive domain*. IP&M, 41 (5), pp. 1101-1119, 2005.

[15] Foley, C., Smeaton, A. F., and Lee, H. *Synchronous collaborative information retrieval with relevance feedback*. 2nd Int. Conf. on Collaborative Computing, pp. 1-4, 2006.