# PIR based on Explicit and Implicit Feedback

Alberto Andreu-Marín, Fernando Javier Martinez-Santiago, Manuel Carlos Díaz-Galiano, and L. Alfonso Ureña-López

Intelligent System for Information Access (SINAI)
Advanced Studies Center in Information and Communication Technologies (CEATIC)
Universidad de Jaén

**Abstract.** Our research aim is twofold: the first one is to get leverage from the relevance feedback that the user provides over the course of every search session. In this way, we explore PIR task as a text categorization problem in order to distinguish between relevant and not relevant documents for the given user. A number of supervised machine learning algorithms has been applied in order to accomplish this task. The second one is related to implicit feedback. More concisely, time spent reading documents and text complexity. From the point of view of text complexity, we propose the hypothesis that there are significant differences of perplexity between judged/non-judged documents by the user. We find some weak statistical evidence that points out that high perplexity and judged documents are correlated.

**Keywords:** Personalised Information Retrieval (PIR) · explicit feedback · implicit feedback · language models · perplexity.



## 1 Introduction

Information Retrieval (IR) is a discipline that involves the retrieval of certain information within a document collection based on specific information needs[12, 2]. The widespread use of IR Systems by a large part of the population, in addition to the huge amount of data generated daily, requires knowledge of the specific information requirements of users. The evaluation that a user makes of a proposed result is personal [6], this can be expressed either explicitly (giving his/her opinion, the so-called relevance judgments) or implicitly (analyzing the behavior while interacting with the system). In addition, it is desirable to obtain a knowledge base about the preferences of each user in order to adapt the proposed results.

The subject of this paper concerns task 2 of the PIR-CLEF laboratory. The challenge in this work is to use the files provided by the organization (csv1 - csv5) to create user profiles that can be used to improve the quality of the results provided by an IR System. Both explicit and implicit data can be found in these files, although certain characteristics can be inferred from the former.

This entire process involves the development of a Personalized Information Retrieval system (PIR) based on a number of explicit and implicit feedbacks

such as user document relevance assessments, search logs including timestamps of every action, and ranking of documents as a result of every search performed by the user.

## 2 Explicit and implicit feedback

Two different types of input data sources are usually identified in the scientific literature: implicit and explicit feedback [7]. Specifically, explicit feedback refers to a conscious assessment given by the user indicating the relevance of a document retrieved for a query. On the other hand, in the implicit feedback, the information must be inferred according to the data that could be collected about the user's behavior when interacting with the system (navigation logs, task description, eyes tracking, etc.).

When implicit feedback is used for the development of this type of system, there may be uncertainty when interpreting the results. To this end, an attempt will be made to find some kind of correlation between the implicitly inferred information and the explicit evaluation given by the user.

## 3 Experiments and results

In this section, we present the different actions that we have accomplished in our participation in CLEF PIR 2018 Task 2 Evaluation of Personalised Information Retrieval.

### 3.1 Feature extraction from retrieved web pages

Firstly, data acquisition has been accomplished to retrieve the Clueweb documents that each user evaluated as a result of the execution of each of them [11]. For this first step, the attribute "query_text" belonging to the file `csv2.csv` provided by the organization is used. A query result is made of the first one hundred documents obtained using the API provided by the organisation and developed by the University of Dublin[1], each of which contains an attribute called "id", which is used to download the corresponding web page using page rendering of Clueweb12 data-set online services[2].

Secondly, once the corresponding web pages are downloaded we move on to the pre-processing phase:

- Only the text contained in the labels "title" and "p" (paragraph) has been considered.
- The resulting text has been processed by eliminating stop words and by executing the Porter stemming algorithm.

[1] http://clueweb.adaptcentre.ie/WebSearcher/search?query="query String"&selection=[selection numbers separated by comma] (last visited: 30/05/2018)

[2] https://www.lemurproject.org/clueweb12/services.php (last visited: 30/05/2018)

- Every document is represented as a set of tokens made by unigrams, bigrams and trigrams.
- TF.IDF has been calculated for each token belonging to each of the different user categories [5]. This determines how relevant this term is in the category.

### 3.2 Explicit feedback as text categorization

Relevance Feedback is a well-known technique in the field of information retrieval. The idea behind relevance feedback is to take the results that are initially returned from a given query, to gather user feedback, and to use information about whether or not those results are relevant to perform a new query [12]. On the other hand, methods based on supervised machine learning are the most frequent approach to accomplish the task of text Categorization [8, 13]. In this section, we face the PIR task as a text categorization task where the categories are the user relevance judgments for every query and a set of retrieved documents. Thus, we do not perform a new query as usual when relevance feedback is applied. Instead of that, we train a number of supervised machine learning algorithms by using features extracted from every document (Section 3.1). The main hypothesis that we want to explore is whether the subjective and non-expert relevance judgments provided for each user are valid to define document categories. A second hypothesis is to validate the set of features extracted from the retrieved documents. A final hypothesis regards with the number of examples (judged documents) provided for each user and query: is this number high enough to train some of the most popular supervised machine learning algorithms?

As a consequence, a number of supervised machine learning algorithms have been trained using the user relevance judgments, characterizing each of the documents based on sequences of n_grams (Unigrams, Bigrams and Trigrams) extracted from the texts and using the bag of words method, which allows us to represent documents ignoring the order of the words that make it up.

**3.2.1 Results.** In order to evaluate our approach, we have followed a k-fold cross-validation approach with $k = 10$ [9] where applicable. In addition we have implemented both a fine-grained and a coarse-grained text categorization task. For the first one, on one hand, we distinguish four categories in the same way the relevance of the document to the topic (1 off-topic, 2 not relevant, 3 somewhat relevant, 4 relevant). On the other hand, The coarse-grained text categorization task is a binary classifier with two categories only: relevant (categories 3 and 4) and non-relevant documents (categories 1 and 2).

The results obtained in Table 1 represent the success rates of the classifiers in the test phase. These data must be analyzed from the perspective of the problems presented by the classification algorithms used. In this case, not all users present sufficient or sufficiently good data to confirm the results presented.

For experiments with 4 categories (1-2-3 and 4), 10-fold cross-validation can only be accomplished for user_02, user_11, user_15 and user_07 in the Books

**Table 1.** Classification Results. Naive Bayes, Support Vector Machine and K-nn classifiers were used.

| | 4 categories | | | 2 categories | | |
|---|---|---|---|---|---|---|
| **Users/topic** | **NB** | **MVC** | **Knn** | **NB** | **MVC** | **Knn** |
| User_02/Music | 0.50 | 0.42 | 0.40 | 0.63 | 0.66 | 0.47 |
| User_07/Books | 0.43 | 0.44 | 0.31 | 0.53 | 0.58 | 0.62 |
| User_07/Travel | 0.66 | 0.75 | 0.71 | 0.79 | 0.83 | 0.73 |
| User_08/Books | 0.79 | 0.73 | 0.75 | 0.91 | 0.87 | 0.71 |
| User_08/Travel | 0.68 | 0.67 | 0.88 | 0.92 | 0.89 | 0.77 |
| User_11/Travel | 0.46 | 0.47 | 0.34 | 0.58 | 0.63 | 0.71 |
| User_12/Sports | 0.47 | 0.42 | 0.33 | 0.79 | 0.74 | 0.38 |
| User_12/Travel | 0.64 | 0.64 | 0.72 | 0.82 | 0.77 | 0.88 |
| User_13/Sports | 0.68 | 0.68 | 0.85 | 0.81 | 0.86 | 0.70 |
| User_15/Travel | 0.47 | 0.46 | 0.56 | 0.69 | 0.74 | 0.65 |
| User_16/Travel | 0.59 | 0.56 | 0.64 | 0.72 | 0.66 | 0.56 |
| User_17/Books | 0.75 | 0.66 | 0.55 | 0.96 | 0.97 | 0.86 |
| User_18/Books | 0.31 | 0.15 | 0.58 | 0.68 | 0.68 | 0.56 |
| User_18/Travel | 0.42 | 0.47 | 0.25 | 0.47 | 0.55 | 0.25 |

| NB & MVC | |
|---|---|
| ■ | 10 fold |
| ■ | 9 fold |
| ■ | 8 fold |
| ■ | 6 fold |
| ■ | 5 fold |
| ■ | 3 fold |
| ■ | 1 fold |

| KNN | |
|---|---|
| ■ | Optimal K |
| ■ | k=3 |

topic only. This is because many of the samples provided by the organization are not representative enough to perform cross-validation up to 10 folds. In general, it can be said that, when it is possible to define 10-fold, explicit feedback gets systematically results whose performance improvement (47% on average) is statistically significant when considering random values as case base. For the rest of the users, the results presented have been obtained by applying a smaller number of folds because some of the samples are unbalanced.

In the case of coarse-grained experiments, with 2 categories, it has not been possible to balance the evaluation values in any of the cases, the reference values taken to check the result of the classifiers remain the same as in the case of 4 categories. It can be seen that all cases the success rates of the classifiers increase by around 15% with respect to fine-grained results.

The best result is obtained by user 17 and Books as topic. The reason is that this case is extremely unbalanced: 128 of a total of 133 documents are judged as off-topic or not relevant.

Regarding the best classifier, it is not possible to select the best one since this depends on the user and topic. This suggests that a voting strategy could be a good option as future work.

### 3.3 Implicit feedback

Consider our first approximation to Personal Information Retrieval, we have focused on two different issues regarding implicit feedback [7], where various search logs are collected and analyzed to infer attitudes with the aim to assess the relevance of certain items indirectly through a users actions and behaviours.

More concisely we are focused in two main hypothesis based on timing and statistics language models.

**3.3.1 The time spent on a page.** Based on the information in the data set provided by the organization, it is intended to make use of the time spent by each user evaluating each document. This experiment aims to demonstrate the hypothesis that the time the user spends evaluating a result is related to the interest generated by the content of the website. This is a quite controversial issue finding evidence that supports this hypothesis [3] or not at all [10].

Suddenly, there is no way to be sure about the time spent on a page just by using the search logs provided by the PIR-CLEF organization since there is no timestamps about close document events. Even though this drawback we have tried to accomplish a study to correlate timing and user assessments. Consequently, we interpreted the time spent on a page as the difference between two consecutive open document events, start session and open document event or open document and end session event. The next step we tested is whether this sequence of time periods follows a normal sequence. Thus, we apply a ShapiroWilk test for each user search log. In addition, we eliminated time periods that are out the two central percentiles 25, with the aim of removing anomalous time periods. We find that the most of time distributions are not normal even though filtering anomalous data. Finally, we applied a two tailed Student T-Test for those cases where normality is found. For the rest of cases, we applied Mann-Whitney-Wilcoxon.

In the end of this process we have to conclude that, in general and following the procedure described above, it is not possible to reject the null hypothesis: user document judgments and time spent on a page are correlated.

**3.3.2 Investigating the relationship between language model perplexity and user relevance measures.** The canonical measure of goodness of a statistical language model is normally reported in terms of perplexity: the exponential of the negative normalized predictive likelihood under the model, and gives an indication of the expected word error rate as in speech recognizers [1]. This finds some evidence that the perplexity of the language model has a systematic relationship with the achievable precision recall performance by using traditional Information Retrieval systems. Following this finding, we hypothesize that for a given probabilistic language model, there is significant differences between the perplexity of the set of documents that are evaluated by the user and those documents that are not evaluated.

We used trigram language models with interpolated Kneser-Kney discounting trained using the SRI language modeling toolkit [14]. We generated different models by varying the training corpus.

– Simple-wiki [4]: 137K sentence Simple English Wikipedia articles.

– Sphinx-70k: CMUSphinx US English generic acoustic model[3]. this is the most general language model that we have considered. This is the best suited to represent the English language.
– ClueWeb12 search: List of documents retrieved by using every set of queries related for each topic. ClueWeb12 search service provided by the organization was applied in order to retrieve the 100 first ranked documents. Note that we have a different language model for each topic proposed in the PIR-CLEF dataset.

The perplexity of three sets of documents per each user query was measured: the set of relevant documents (user relevance judgment is 3 or 4), non-relevant documents (user relevance judgment is 1 or 2) and unjudged documents (there is no user relevance judgment in spite of they are part of the ranked list of documents retrieved)

Finally, an one-tailed Mann-Whitney-Wilcoxon (MWW) test was performed[4]. When language models based on Simple-wiki and ClueWeb12 search datasets are applied we have no found any significative difference between the perplexity of the the three set of documents considered (relevant, non-relevant or unjudged) When Sphinkx-70k is used to train the language model, we find some evidence that the complexity of judged documents (relevant or not relevant) are greater than those that are unjudged (U-value=59, critical U-value at p<0,05=51). This is quite surprising since it could be interpreted in the way that the user tends to evaluate the most complex texts. Once we revise some of the non-judged document we find that it is quite frequent that these documents do not have textual content at all, only lists of sections, menus and stylesheets but no or very little meaningful text.

## 4 Conclusion and Future Work

In this work, an overview has been given of the use of explicit and implicit feedback for the generation of user profiles. In the case of explicit feedback, a classification system is trained based on the user's judgements of relevance provided for a given set of documents. In the case of implicit feedback, the intention is to seek a correlation between the information that can be inferred from the data and the relevance judgments provided by users, so that user profiles and system accuracy can be improved.

In future work it would be interesting to have information relevant to certain aspects. The file `csv6.csv` provided by the organization, shows the TF.IDF

[3] https://sourceforge.net/projects/cmusphinx/files/Acoustic and Language Models/US

[4] When the dataset is small, the P-Value from t-Student is likely to be the most usual test but it requires a normal distribution of the dataset. For this reason, we applied Shapiro-Wilk test that is suited for small datasets and we found that it is not always possible to assert that the considered datasets follow a normal distribution. As a consequence we applied a non-parametric test, the Mann-Whitney-Wilcoxon U test

values of the tokenized terms, it would be interesting to be able to relate each of these terms with the document to which it belongs, in this way, they could be used for the realization of the training phase characterization of web pages.

On the other hand, in order to search for a correlation between the interest that an user has in a web page and the time he spends visiting it, it would be interesting to have the "CLOSE_DOCUMENT" attribute of all the records. Currently there are only 5 records with this attribute in the `csv2.csv` file.

In order to further deepen this task, it would be helpful to have the traditional expert judgements of relevance. In this way, relationships could be sought in those cases in which the user did not agree with the expert and try to discern the most probable cause.

## 5 Acknowledgments

Work supported by a grant from the Ministry of Education, Culture and Sport (MECD-Scholarship BES-2016-076609) and the REDES project (TIN2015-65136-C2-1-R) of the Spanish Government.

## References

1. Azzopardi, L., Girolami, M., van Risjbergen, K.: Investigating the relationship between language model perplexity and ir precision-recall measures. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. pp. 369–370. ACM (2003)
2. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463. ACM press New York (1999)
3. Claypool, M., Le, P., Wased, M., Brown, D.: Implicit interest indicators. In: Proceedings of the 6th international conference on Intelligent user interfaces. pp. 33–40. ACM (2001)
4. Coster, W., Kauchak, D.: Simple english wikipedia: a new text simplification task. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. pp. 665–669. Association for Computational Linguistics (2011)
5. Galarza Quishpe, E.D.: Text Classification for literature search of research study designs. Master's thesis, Australia/Universidad de Melbourne (2015)
6. Janes, J.W.: Other people's judgments: A comparison of users' and others' judgments of document relevance, topicality, and utility. Journal of the American Society for Information science **45**(3), 160 (1994)
7. Jannach, D., Lerche, L., Zanker, M.: Recommending based on implicit feedback. In: Social Information Access, pp. 510–569. Springer (2018)
8. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: European conference on machine learning. pp. 137–142. Springer (1998)
9. Kohavi, R., et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai. vol. 14, pp. 1137–1145. Montreal, Canada (1995)
10. Morita, M., Shinoda, Y.: Information filtering based on user behavior analysis and best match text retrieval. In: SIGIR94. pp. 272–281. Springer (1994)

11. Pasi, G., Jones, G.J., Marrara, S., Sanvitto, C., Ganguly, D., Sen, P.: Evaluation of personalised information retrieval at clef 2017 (pir-clef): towards a reproducible evaluation framework for pir
12. Salton, G., Buckley, C.: Improving retrieval performance by relevance feedback. Journal of the American society for information science **41**(4), 288–297 (1990)
13. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) **34**(1), 1–47 (2002)
14. Stolcke, A.: Srilm-an extensible language modeling toolkit. In: Seventh international conference on spoken language processing (2002)