An Interactive Two-Dimensional Approach to Query Aspects Rewriting in Systematic Reviews. IMS Unipd At CLEF eHealth Task 2.

Giorgio Maria Di Nunzio[1], Federica Beghini[2], Federica Vezzani[2], and Geneviève Henrot[2]

[1] Dept. of Information Engineering – University of Padua
[2] Dept. of Linguistic and Literary Studies – University of Padua
giorgiomaria.dinunzio@unipd.it, fede.beghini92@gmail.com, federicavezzani92@gmail.com, genevieve.henrot@unipd.it

Abstract. In this paper, we describe the participation of the Information Management Systems (IMS) group at CLEF eHealth 2017 Task 2. This task focuses on the problem of systematic reviews, that is articles that summarise all evidence that is published regarding a certain medical topic. This task, known in Information Retrieval as the total recall problem, requires long and tedious search sessions by experts in the field of medicine. Automatic (or semi-automatic) approaches are essential to support these type of searches when the amount of data exceed the limits of users, i.e. in terms of attention or patience. We present the two-dimensional probabilistic version of BM25 with explicit relevance feedback together with a query aspect rewriting approach for both the simple evaluation and the cost-effective evaluation.

1 Introduction

In this paper, we describe the participation of the Information Management Systems (IMS) group at CLEF eHealth 2017 [1] Task [2]. This task focuses on the problem of systematic reviews, that is articles that summarise all evidence that is published regarding a certain medical topic. This task, known in Information Retrieval as the total recall problem, requires long and tedious search sessions by experts in the field of medicine. Automatic (or semi-automatic) approaches are essential to support these type of searches when the amount of data exceed the limits of users, i.e. in terms of attention or patience. In particular, the aim is to make research papers abstract and title screening more effective given the results of a boolean search submitted to a medical database.

The CLEF eHealth Task 2 has two types of evaluation procedures to assess the quality of a system that supports systematic reviews. These procedures are based on the following assumptions:

– Simple evaluation, the user of the system is the researcher (end-user) that performs the abstract and title screening of the retrieved articles. Every time the system returns an abstract to the end-user there is an incurred cost.

– Cost-effective evaluation, the user that performs the screening is not the end-user. The user can interchangeably perform abstract and title screening, or document screening, and decide what documents to pass to the end-user. Every time the system provides an abstract to the user, she/he can i) either read the abstract (with an incurred cost, like in the simple evaluation) and decide whether to pass this document to the end-user, ii) or read the full document (with a higher cost) and decide whether to pass this document to the end-user, iii) or directly pass the document to the end-user. For each document passed to the end-user there are additional costs that need to be added.

The objective of our participation to this task was to:

– find the best parameters (in terms of classification/ranking accuracy) of the BM25 model [4];
– explore the problem of query aspects and query (re-)formulation given an information need [6, 10];
– integrate the query aspects into the two-dimensional probabilistic model [3];
– study an automatic feedback loop to find the optimal stopping strategy [8].

2 Approach

In this paper, we continue to investigate the interaction with the two dimensional interpretation of the BM25 model applied to the problem of explicit relevance feedback with three goals in mind:

– we want to create a set of relevance judgements with the least effort by human assessors,
– we use interactive visualizations to interpret the intermediate results of the relevance feedback,
– we use explicit query rewriting by non experts to create different aspects of the information need.

Following the work started in [6, 4, 8, 3, 7], we continue to study the two-dimensional interpretation of the BM25 model defined in the following section.

2.1 BM25

The BM25 is a probabilistic retrieval model where, if we use the definition given by Zaragoza and Robertson in [9], the weight of the i-th term in a document is equal to:

$$w_i^{BM25}(tf) = \frac{tf}{k_1\left((1-b)+b\frac{dl}{avdl}\right)+tf} w_i^{BIM} \tag{1}$$

where k_1 and b are two parameters (we used the default values used by Terrier[3], $k_1 = 1.2$ and $b = 0.75$), tf is the term frequency in the document, and w_i^{BIM} is

[3] `http://terrier.org`

the Binary Independence Model weight of the i-th term:

$$w_i^{BIM} = \log \frac{\theta_i^{\mathcal{R}}}{(1-\theta_i^{\mathcal{R}})} \frac{(1-\theta_i^{\mathcal{NR}})}{\theta_i^{\mathcal{NR}}} \tag{2}$$

where $\theta_i^{\mathcal{R}}$ and $\theta_i^{\mathcal{NR}}$ are the parameters of the Bernoulli random variable that represent the presence (or absence) of the i-th term in the relevant ($\mathcal{R}$) and non-relevant ($\mathcal{NR}$) documents. The estimate of each parameter is:

$$\theta_i^{\mathcal{R}} = \frac{r_i + \alpha^{\mathcal{R}}}{R + \alpha^{\mathcal{R}} + \beta^{\mathcal{R}}} \tag{3}$$

$$\theta_i^{\mathcal{NR}} = \frac{n_i - r_i + \alpha^{\mathcal{NR}}}{N - R + \alpha^{\mathcal{NR}} + \beta^{\mathcal{NR}}} \tag{4}$$

where R is the number of relevant documents, r_i the number of relevant documents in which the i-th term appears, N is the total number of documents and n_i is the total number of documents in which the i-th term appears. Parameters α and β correspond to the hyper-parameter of the conjugate beta prior distribution of the Bernoulli random variable. For $\alpha^{\mathcal{R}} = \beta^{\mathcal{R}} = 0.5$ and $\beta^{\mathcal{R}} =^{\mathcal{NR}}= 0.5$, we obtain the definition of the well-known Robertson - Spärck Jones weight w_i^{RSJ} [9].

2.2 Two-Dimensional Model

The two-dimensional representation of probabilities [5, 8] is an intuitive way of presenting a two-class classification problem on a two-dimensional space. Given two classes, for example relvant $\mathcal{R}$ and non-relevant $\mathcal{NR}$, a document d is assigned to category $\mathcal{R}$ if the following inequality holds:

$$\underbrace{P(d|\mathcal{NR})}_{y} < m \underbrace{P(d|\mathcal{R})}_{x} + q \tag{5}$$

where $P(d|\mathcal{R})$ and $P(d|\mathcal{NR})$ are the likelihoods of the object d given the two categories, while m and q are two parameters that can be assigned (automatically or by a user) to compensate for either the unbalanced class issues or different misclassification costs.

If we interpret the two likelihoods as two coordinates x and y of a two dimensional space, the problem of classification can be studied on a two-dimensional plot. The decision of the classification is represented by the line $y = mx + q$ that splits the plane into two parts: all the points that fall 'below' this line are classified as objects that belong to class $\mathcal{R}$.

Two-dimensional BM25 In order to link the two-dimensional model to the BM25 model, first we define the BIM weight as a difference of logarithms:

$$w_i^{BIM} = \log \frac{\theta_i^{\mathcal{R}}}{(1-\theta_i^{\mathcal{R}})} - \log \frac{\theta_i^{\mathcal{NR}}}{(1-\theta_i^{\mathcal{NR}})} = w_i^{BIM,\mathcal{R}} - w_i^{BIM,\mathcal{NR}} \tag{6}$$

then, we can define the BM25 term weight accordingly

$$w_i^{BM25}(tf) = \frac{tf}{k_1\left((1-b)+b\frac{dl}{avdl}\right)+tf}\left(\ w_i^{BIM,R} -\ w_i^{BIM,NR}\right) \tag{7}$$

We now have all the elements to define the two coordinates $x = P(d|\mathcal{R})$ and $y = P(d|\mathcal{NR})$ in the following way:

$$P(d|\mathcal{R}) = \sum_{i\in d} w_i^{BM25,\mathcal{R}}(tf) \tag{8}$$

$$P(d|\mathcal{NR}) = \sum_{i\in d} w_i^{BM25,\mathcal{NR}}(tf) \tag{9}$$

where $\sum_{i\in d}$ indicates (with an abuse of notation) the sum over all the terms of document d.

3 Method

Given the definition of two-dimensional BM25 model, we focused on the following problems:

1. find the best combination of hyper-parameters $\alpha^{\mathcal{R}}$, $\alpha^{\mathcal{NR}}$, $\beta^{\mathcal{R}}$, $\beta^{\mathcal{NR}}$,
2. devise a strategy to create different query aspects of the same information need given a minimum amount of relevance feedback,
3. explore different options of explicit relevance feedback for both the simple and the cost-effective evaluation subtasks.

For step 1), we used the training data available for this task to find the best combination of parameters trough a force brute approach.

For step 2), we decided to use the following procedure:

- for each topic, we run a plain BM25 retrieval model and get the relevance feedback for the first abstract in the ranking list,
- we get the explicit relevance feedback on that abstract and ask to two different people (non-experts in the field of medicine) to review the abstract and rewrite an alternative query also according to the value of the feedback (relevant or not),

For example, for topic CD008803 the original information need is expressed by the following statement:

"Optic nerve head and fibre layer imaging for diagnosing glaucoma"

we run BM25 and obtain the top retrieved abstract is document 19028735, the content of which is:

> title: Imaging of the retinal nerve fibre layer for glaucoma.
> abstract: Glaucoma is a group of diseases characterised by retinal ganglion cell dysfunction and death. Detection of glaucoma and its progression are based on identification of abnormalities or changes in the optic nerve head (ONH) or the retinal nerve fibre layer (RNFL), either functional or structural. This review will focus on the identification of structural abnormalities in the RNFL associated with glaucoma. ...

Then we pass the information that this abstract is not relevant (according to the abstract qrels) to the two users that rewrite the query, and we obtain two new query aspects. One user was asked to prepare a list of terms:

> "optic nerve head, ONH, optic disc, fibre layer, diagnosis, retinal, imaging, RNFL, glaucoma, SLP, Scanning laser polarimetry, HRT, Heidelberg Retina Tomograph, OCT, Optical Coherence Tomography, GDx"

The other user had to write a sort of information need instead of a list of words:

> "Diagnostic accuracy of HRT, OCT and GDx for diagnosing manifest glaucoma by detecting ONH and RNFL damage."

The first type of query was written with the aim of entering the key words contained in the topic title, in the boolean query and in the article that was given (if relevant), along with other terms which were the result of various processes: the componential analysis of some words, the variants, the synonyms, the declensions and the acronyms of some terms and the reading of other relevant information using sources on the web[4]. The componential analysis consists of breaking down the sememe (i.e. the meaning) of a word in all its sense components[5], e.g. the semes of radiculopathy[6](topic CD007431) are /pathology/, /nerve root/, /spinal/, /inflammation/, /compression/. Therefore, in this case we also included all these terms in the query, which were not present in the information need[7]. We did not decide to enter the semes of all the words, but just of the terms whose semes we thought could improve the search of the most

[4] PubMed `https://www.ncbi.nlm.nih.gov/pubmed/`
The Free Dictionary by Farlex - Medical Dictionary `http://medical-dictionary.thefreedictionary.com/radiculopathy`
Merriam Webster Dictionary `https://www.merriam-webster.com/)`
Wikipedia `https://en.wikipedia.org/wiki/Main_Page`

[5] Rastier, F, (1987), Smantique interprtative, d. Presses Universitaires de France, 2009, Paris, p.18-32.
Dubois., J. et al. (1994), Dictionnaire de linguistique et des sciences du langage, d.Larousse, Paris, p.423-424.
Ducrot, O., Schaeffer, J.-M., (1972), Nouveau dictionnaire encyclopdique des sciences du langage, d.du Seuil, 1995, Paris, p.445-447.

[6] The Free Dictionary by Farlex - Medical Dictionary `http://medical-dictionary.thefreedictionary.com/radiculopathy`

[7] Physical examination for lumbar radiculopathy due to disc herniation in patients with low-back pain.

relevant articles. Furthermore, if the terms had many variants, we added them to the query: e.g. in topic CD008760, we did not just enter oesophageal and oesophagus, but also esophageal and esophagus. Moreover, we tried to use not only one grammatical form to describe a concept, which is why we did not just enter nouns, but also verbs and adjectives, e.g. radiculopathy, radicular and spinal, spine (topic CD007431); endometriosis, endometrial (topic CD012019), diagnosis, diagnose, diagnosing, diagnostic (topic CD010542). We also added synonyms, e.g. diagnosis, screening, examination (topic CD009925) and diagnosis, detection (topic CD010783). For what concerns the process of declension, sometimes we wrote not only the singular, but also the plural form of a noun, e.g. dementia, dementias; biomarker, biomarkers (topic CD008782). Then, we entered the acronym of some terms, e.g. LPB (lumbago) (topic CD007431); mild cognitive impairment (MCI) (CD008782). Finally, the terms have been entered in a random order.

The second type of query was written with the aim of reformulating the information need. The purpose was to rewrite the information given for each topic using an alternative terminology and trying to reformulate a meaningful and humanly readable sentence. For this reason, validly attested synonyms and orthographic alternatives were used as variants of the medical terms provided in the original information need. In addition, another criterion was to systematically replace acronyms with their expansions and expansions with their acronyms. For example, for topic CD009135, the information need "Rapid tests for the diagnosis of visceral leishmaniasis in patients with suspected disease" was rewritten using synonyms and acronyms for "visceral leishmaniasis": "Evaluation of rapid examinations in order to detect VL (kala-azar, black fever and Dumdum fever) in patients with clinically suspected infection". This approach allowed us to expand the medical terminology and to evaluate also the documents in which the selected variants were present. The sources from which the terminological variants were selected were PubMed, the online medical dictionary Merriam Webster and Wikipedia. For what concerns the topics presenting a relevant document (relevance index 1) selected by the expert, the criterion of re-writing the information need was based on the knowledge acquired by reading the PubMed article abstract. This step facilitated the reformulation of the title by focusing on the typology of the request and its related aspects. On the contrary, the topics where the document's relevance index was 0, the reformulation was based on the terminology frequency analysis and on an in-depth research of the topic on reliable sources available on the web.

For step 3), we designed alternative strategies that use the following parameters:

- number of documents to assess, in batches or iteratively,
- percent of documents to assess,
- maximum number of documents to assess per iteration,
- number of terms to add at each feedback iteration,
- for the cost-effective evaluation, the minimum precision the system can reach before stopping the search.

Simple evaluation For the simple evaluation subtask, we focused on the number (or percentage) of documents to use for explicit relevance feedback and how to combine the query aspects. No threshold on the number of documents to retrieve was set.

Cost-effective evaluation For the cost-effective subtask, we performed two rounds of relevance feedback: first retrieve, then classify. In the first round, we select a percentage of documents for explicit relevance feedback; then, we use the relevance information to build the two classes $\mathcal{R}$ and $\mathcal{NR}$. Once the two classes are built, we use the two-dimensional space to pick the document with partial recall 100% (by 'partial', we mean that if during the iteration we retrieve 10 relevant document out of 20, we pick the relevant document with the lowest score) and let the classification line pass through that point. Then we iterate the feedback until precision reaches 0.2.

In Figure 1, we show the two dimensional situation at four different steps of the iteration. Green dots represents relevant documents, red dots non-relevant documents, black dots documents to be ranked (or judged). In Figure 1 (a), we see the documents at the end of the relevance feedback phase. After we re-set the probabilities by building the two classes of relevant and non relevant documents, the documents are in a different position of the two-dimensional space, Figure 1 (b). The space between the interpolating line of the relevant documents (dashed line) and the line of the last relevant document (dot-dashed line) is the 'grey area' where we expect to find more relevant documents. After a few iteration, the relevant and non relevant clouds of points become more and more separate, Figure 1 (c). When all the documents within the space between the two lines are judged (plus some other of the 'extra-rounds') the systems stops sending documents to the user, Figure 1 (d).

4 Experiments

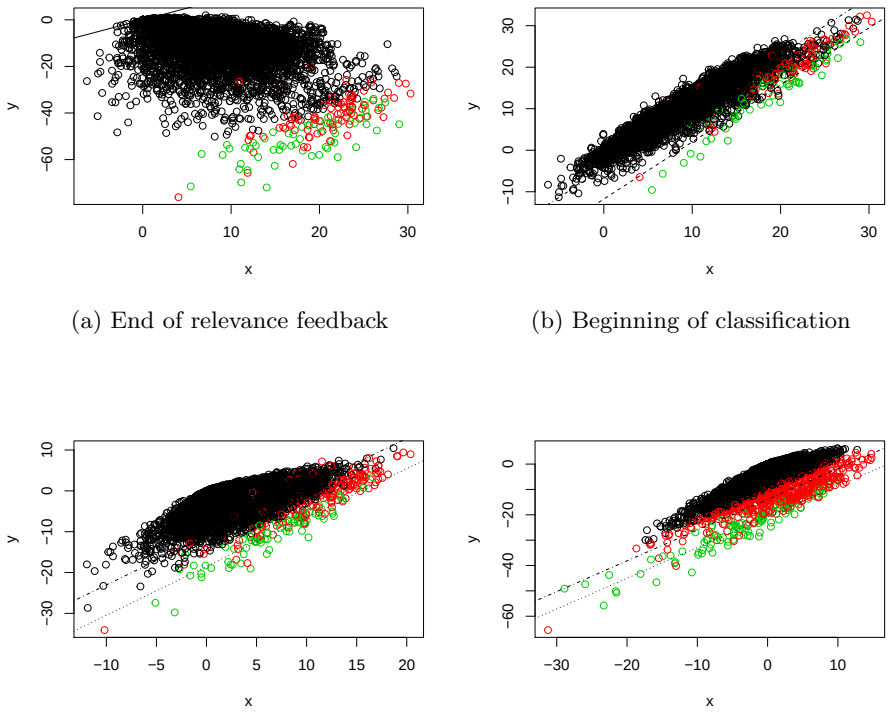
In all experiments, we used the first document retrieved with a BM25 approach (and then judged) to build two different queries that represent the same information need. The two alternative queries are combined with the original one in different ways as described in the following sections.

For all the experiments, we set the best set of values for the parameters $\alpha^{\mathcal{R}}$, $\alpha^{\mathcal{NR}}$, $\beta^{\mathcal{R}}$, $\beta^{\mathcal{NR}}$ of the BM25 found with a brute force approach on the training data. The values are:

- $\alpha^{\mathcal{R}} = \alpha^{\mathcal{NR}} = 1.0$
- $\beta^{\mathcal{R}} = \beta^{\mathcal{NR}} = 0.01$

These values are consistent with other experiments and indicate that a beta prior distribution that discounts the 'presence' of a term in favour of its 'absence' (high α and low β) results in a better retrieval performance.

We also run a set of experiments on the training data to find the value of the number of documents k to use for relevance feedback that gives the best



(a) End of relevance feedback

(b) Beginning of classification

(c) Second round of classification

(d) End of classification

Fig. 1: Cost-effective approach on the two-dimensional space. Green dots represent relevant documents, red dots non-relevant documents, black dots documents to be ranked (or judged). The dashed line shows the interpolating line of the relevant documents, while the dot-dashed line indicates the last relevant documents found. When all the documents within this space are judged (plus some other of the 'extra-rounds') the systems stops sending documents to the user.

trade-off between cost and effectiveness, and we found that $k = 50$ is a good estimate.

4.1 Simple Evaluation

For the simple evaluation subtask, we submitted four runs:

- **ims_iafa_m10k150f0m10**, run-1, this run uses Interactive Automatic Feedback with query Aspects (iafa) and, for each topic, uses k = 150 feedback rounds where, at each round, a new word is picked from the relevant documents and the top document is judged. For each topic, a total of 150 documents are judged.
- **ims_iafas_m10k50f0m10**, run-2, this run uses Interactive Automatic Feedback with query Aspects with Separate rankings (iafas). At each round of feedback, the two query variants are run in parallel with the original one and three different documents are judged. There are k = 50 rounds for a total of 150 documents judged per topic.
- **ims_iafap_m10p2f0m10**, run-3, this run uses Interactive Automatic Feedback with query Aspects using a Percent (iafap) of documents for feedback. This run is similar to the first one but it uses a number of documents for relevance feedback that is proportional to the number of documents to rank. In this case, $p = 2$ is two percent of feedback.
- **ims_iafap_m10p5f0m10**, run-4, this run uses Interactive Automatic Feedback with query Aspects using a Percent (iafap) of documents for feedback. The percent of feedback is $p = 5$.

4.2 Cost-Effective Evaluation

For the cost-effective evaluation subtask, we submitted four runs. All the four runs use the same approach named Interactive Automatic Feedback with query Aspects and Percent of relevance feedback and Classification (iafapc). In particular, we tried different values of parameters concerning the percent of documents for relevance feedback and the maximum number of documents for relevance feedback in the initial phase.

During the classification phase, we calculate the linear interpolation of relevant documents if 5 or more relevant documents are available, otherwise we compute the linear interpolation of non relevant documents. If the angular coefficient of the line is less than 0.9, we adjust it. We iterate this process by selecting the top 10 documents and perform explicit relevance feedback until precision reaches 0.2. After that point, extra iterations are performed with half of the documents used in the previous feedback round. We stop if no other documents are available or precision is below 0.2 and we have only one document for the extra rounds of relevance feecback.

- **ims_iafapc_m10p5f0t0p2m10**, run-5, this run uses 5 percent of relevance feedback documents per round of relevance feedback in the initial phase.

Table 1: Simple evaluation results. Top part shows abstract qrels evaluation, bottom part document qrels evaluation.

run	ap	last_rel	wss100	wss95	norm_area	total_cost	loss_er	loss_r
run-1	.280	2269.333	.415	.508	.896	4075.233	.544	.000
run-2	.266	2304.600	.410	.517	.892	4206.567	.544	.000
run-3	.253	2395.533	.366	.476	.875	4076.367	.544	.000
run-4	.269	2260.467	.398	.496	.885	4311.433	.544	.000
run-1	.223	1055.793	.706	.713	.932	3935.414	.544	.000
run-2	.190	990.000	.706	.723	.928	4065.345	.544	.000
run-3	.202	838.897	.661	.685	.919	4156.517	.544	.000
run-4	.212	1007.379	.706	.703	.931	4311.433	.544	.000

Table 2: Cost-effective evaluation results. Top part shows abstract qrels evaluation, bottom part document qrels evaluation.

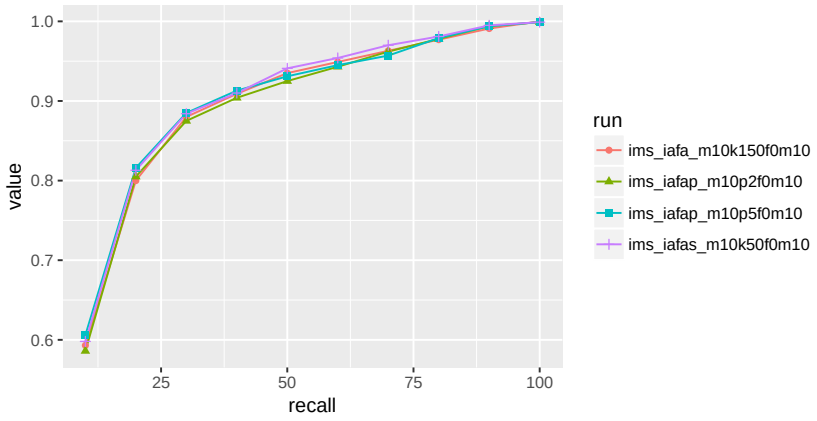
run	ap	last_rel	wss100	wss95	norm_area	total_cost	total_uni	total_wei	loss_er	loss_r
run-5	.232	540.400	.151	.176	.755	653.033	1488.064	5021.993	.115	.097
run-6	.244	379.533	.133	.168	.774	478.300	2228.266	6008.652	.124	.053
run-7	.264	396.800	.183	.247	.808	511.633	2238.965	5969.683	.161	.044
run-8	.270	615.933	.255	.411	.859	807.300	1714.993	4813.495	.169	.019
run-5	.190	364.034	.428	.529	.845	633.931	1029.042	2292.358	.091	.063
run-6	.200	280.345	.385	.441	.856	474.966	1365.765	3813.256	.119	.023
run-7	.216	280.379	.480	.535	.886	509.448	1388.187	3833.256	.170	.017
run-8	.217	414.586	.578	.638	.913	793.69	1310.866	2615.199	.206	.007

- **ims_iafapc_m10p10f0t150p2m10**, run-6, this run uses 10 percent of relevance feedback and a maximum of 150 documents per round of relevance feedback in the initial phase.
- **ims_iafapc_m10p20f0t150p2m10**, run-7, this run uses 20 percent of relevance feedback and a maximum of 150 documents per round of relevance feedback in the initial phase.
- **ims_iafapc_m10p20f0t300p2m10**, run-8, this run uses 20 percent of relevance feedback and a maximum of 300 documents per round of relevance feedback in the initial phase.

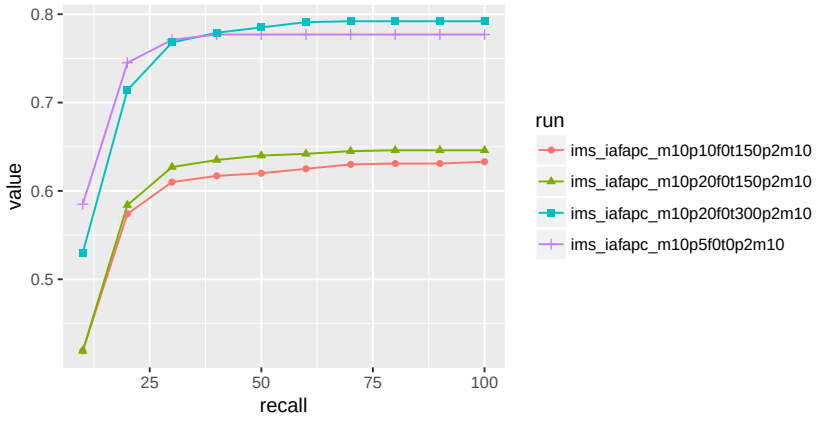
The results for the simple evaluation are reported in Table 1 and Figure 2a while the results for the cost-effective evaluation are reported in Table 2 and Figure 2b.

5 Final Remarks and Future Work

In this paper, we presented the experiments of our research group to the CLEF eHealth Task 2. The objective of our participation to this task was to investi-



(a) NCG for simple evaluation



(b) NCG for cost-effective evaluation

Fig. 2: NCG at different recall values for the simple and cost-effective evaluation.

gate a better set of parameters for the BM25, explore the problem of query aspects and query (re-)formulation given an information need, integrate the query aspects into the two-dimensional probabilistic model, and study an automatic feedback loop to find the optimal stopping strategy.

Some interesting findings during the training phase that we will document more deeply in the future can be summarised as follows:

- there are values for the α and β parameter that clearly outperform the standard BM25 with $\alpha = \beta = 0.5$;
- performing an iterative explicit relevance feedback one document at a time changes significantly the performance of both retrieval and classification (the cost of training at each round of feedback is very high in computational terms, though);
- adding query aspects to the original information need increase consistently the performance of both the retrieval and classification;
- choosing the right terms to add during the iteration of relevance feedback may change significantly the results of both the retrieval and classification.

The results of the test phase presented in the previous section will be analyzed more deeply in the next weeks. In particular, it is not clear whether a fixed amount of documents ($k = 150$, for example) may be better than a fixed percentage of documents (say, $p = 5$). It will be interesting to study and compare the simple and the cost-effective strategies in terms of the actual costs, as shown by Table 1 and Table 2. We will also continue to investigate the process of query aspect rewriting by extending it to the case of iteratively rewriting the query aspects according to the shifts of the two-dimensional cloud of points, as shown in Figure 2.

References

1. Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Aurélie Névéol, Aude Robert, Evangelos Kanoulas, Rene Spijker, João Palotti, and Guido Zuccon, editors. *CLEF 2017 eHealth Evaluation Lab Overview. CLEF 2017 - 8th Conference and Labs of the Evaluation Forum*, Lecture Notes in Computer Science. Springer, 2017.
2. Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker, editors. *CLEF 2017 Technologically Assisted Reviews in Empirical Medicine Overview. In Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017.*, CEUR Workshop Proceedings. CEUR-WS.org, 2017.
3. Giorgio Maria Di Nunzio. A new decision to take for cost-sensitive naïve bayes classifiers. *Inf. Process. Manage.*, 50(5):653–674, 2014.
4. Giorgio Maria Di Nunzio. Geometric perspectives of the BM25. In *Proceedings of the 6th Italian Information Retrieval Workshop, Cagliari, Italy, May 25-26, 2015.*, 2015.
5. Giorgio Maria Di Nunzio. Interactive text categorisation: The geometry of likelihood spaces. *Studies in Computational Intelligence*, 668:13–34, 2017.

6. Giorgio Maria Di Nunzio, Maria Maistro, and Daniel Zilio. Gamification for IR: the query aspects game. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*, 2016.
7. Giorgio Maria Di Nunzio, Maria Maistro, and Daniel Zilio. Gamification for machine learning: The classification game. In *Proceedings of the Third International Workshop on Gamification for Information Retrieval co-located with 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016), Pisa, Italy, July 21, 2016.*, pages 45–52, 2016.
8. Giorgio Maria Di Nunzio, Maria Maistro, and Daniel Zilio. The university of padua (IMS) at TREC 2016 total recall track. In *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016*, 2016.
9. Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.
10. Kazutoshi Umemoto, Takehiro Yamamoto, and Katsumi Tanaka. Scentbar: A query suggestion interface visualizing the amount of missed relevant information for intrinsically diverse search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 405–414, 2016.