LIG at CLEF 2015 SBS Lab

Nawal Ould-Amer[1], Philippe Mulhem[1], and Mathias Géry[2]

[1] Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
CNRS, LIG, F-38000 Grenoble, France
[2] Université de Lyon, F-42023, Saint-Étienne, France,
CNRS, UMR 5516, Laboratoire Hubert Curien, F-42000, Saint-Étienne, France
Université de Saint-Étienne, Jean-Monnet, F-42000, Saint-Étienne, France
{Nawal.Ould-Amer, Philippe.Mulhem}@imag.fr,
mathias.gery@univ-st-etienne.fr

Abstract. This paper describes the work achieved by the MRIM research group of Grenoble, using some data from the LaHC of Saint-Étienne, in a way to test personalized retrieval of books for the Social Book Search Lab of CLEF 2015. Our proposal rely on a biased fusion of content-only retrieval, using BM25F and LGD retrieval models, user non-social profile based on the catalog of the requester, and social profiles using user/user links generated from their catalogs and ratings on books. The official results obtained show a clear positive impact of user profile, and a small positive impact of the social elements we used. Post official results that present non biased fusion scores are also presented.

Keywords: Fusion of scores, user profile, social links

1 Introduction

This paper describes our participation to INEX Social Book Search Suggestion Track challenge. The goal of this challenge is to evaluate approaches for supporting users in searching collections of books based on book metadata and associated user-generated content [4]. The work described here focuses on several aspects of personalized information retrieval that integrates social networks information. Our objectives during the participation was twofold: a) to rely as much as possible on Information Retrieval Systems to handle non-social and social profiles, and b) to provide a simple integration of the three elements (i.e., content, non-social profiles, social profiles) according to linear combination of scores. Relying heavily on existing tested IR tools allows us to focus on experimenting ideas. Proposing simple score fusions allows us to analyze more easily how configurations behave. More precisely, our experiments conducted for SBS 2015 emphasizes on:

- Studying the impact of linear fusions of score compared to classical weighted linear fusions;
- Studying the impact of fusing several content-only results;

– Studying the impact of using a simple user profile as query extension (non-social profile);
– Studying the impact of generated *friend* relations on the quality of the results (social profile).

From the data provided by the SBS 2015 dataset, the following elements were used at one time or another:

– the fields title, summary, content and tags from the documents: all concatenated for unstructured retrieval, and separated for field-based retrieval using BM25F;
– the fields title, mediated_query and narrative from the topics;
– the documents and ratings from the "topic users" (a topic user is the description of the user that asks a query): used to compute "friendship" relationships between users;
– the documents and ratings from the profiles of the non-topics users: used to compute "friendship" relationships between users.

The IR processes were achieved on the Terrier system[3] [5].

The section 2 focuses on the description of the fusion that was exploited: one original point relates to the *biases* that we propose. Section 3 tackles multiple content-only matching for the documents, as we found out that such integration is beneficial. Then, we introduce in section 4 the use of non-social profiles, and we detail how we defined friendship relations between users, using their catalogs and ratings, as well a the way we used the profiles of such friends when processing queries. Additional processing must be achieved on SBS data to get results. We discuss in section 5 some of these elements before depicting in section 6 the official, as well as some non-official, results obtained, before concluding in section 7.

2 Biased linear fusions of scores

Fusing scores of several IR systems is nontrivial problem. In our case, as described in the introduction, we propose to use *biased* linear fusions of scores, as an extension of the "Zero-one" normalization used by Wu, Crestani and Bi in [8].

Let us focus on a fusion of two results lists L_1 and L_2, composed of couples (doc,rsv). To be realistic, L_1 and L_2 are limited to the top n results. Assume that a document d has a score value of $score(d, L_i)$ in L_i, with $i \in \{1, 2\}$, and d is at rank $rank(d, L_i)$ in L_i; that b_i is the bias of L_i; and that h_i is the horizon (a rank position) above which we do not look at the in results list L_i. The normalized score of d in L_i is then:

$$f(d, L_i) = \begin{cases} (1 - \frac{vmax(L_i) - score(d,L_i)}{vmax(L_i) - vmin(L_i)}) + b_i & \text{if } rank(d, L_i) \leq h_i \\ 0 & \text{otherwise} \end{cases}$$

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

with $vmin$ and $vmax$ the minimal and maximal value of scores in a result list.

Compared to the fitting used by [8], our idea is that we allow different search results to fit into different intervals. This is independent of the way different result lists are combined, but a kind of "boosts" that forces the final score values for a result list to be in $\{0\} \cup [b_i, b_i + 1]$ (the value 0 denotes that the document does not occur in the top h_i elements of the list). This boost is independent from the way the scores are fused afterward. If we make a parallel with the general fitting proposed by [8], our proposal allows an independent scaling for each list fused.

Then, the overall fusion computes a weighted average of the normalized scores (COMB-sum from [7]) using a parameter α that denotes the relative importance of L_1 over L_2, and rerank the results according to the new fused scores. Compared to a usual weighted average, the difference here comes mainly from the b_i. Assigning 0 to all b_is leads to a usual weighted average COMB-sum.

We discuss the impact of such biases in the section dedicated to the experiments.

3 Fusion of content-only scores (run LIG_1)

On experiments conducted over the SBS 2014 dataset, we noticed that fusing several content-only runs had a positive impact with a relative nDCG@10 improvement larger than 10%. That is why we propose to fuse one result coming from BM25F [6] run (parameters values taken from [3]) and one result coming from a Log logistic model [2]. Grid optimization of parameters on SBS 2014 data led to the parameters used for the official run tagged **LIG_1**, described in table 1. The fusion score is computed as follows:

$$\begin{aligned} RSV_{LIG\_1}(Q,d) = \alpha_{BM25F}(Score_{BM25F}(Q,d) + b_{BM25F}) \\ + \alpha_{LGD}(Score_{LGD}(Q,d) + b_{LGD}) \end{aligned} \quad (1)$$

where α_{BM25F} and α_{LGD} are the relative importance of BM25 scores and LGD score respectively, b_{BM25F} and b_{LGD} are bias of each results list. The normalized scores use an horizon h at 1000 documents.

Table 1. Parameters for the content-only run **LIG_1**

	α	b	h
$BM25F$	0.4	0.5	1000
LGD	0.6	0.4	1000

4 Personalized IR exploiting profiles

4.1 Non-social user profile (run LIG_2)

What we depict here as "non-social" corresponds to the individual user data. In our case, these data refer to the catalog of the user. We assume that Cat_u denotes the catalog (list of books) of a given user u (from the corpus U of users). To construct a user profile, we take inspiration from Cai and Li who consider each user profile as a vector of tags and use a L_1 normalized term frequency (NTF) to denote the preference degree of user on a tag [1]. Similarly, we describe the profile $Prof_u$ of a user u as a weighted vector based on Cat_u, where each term is weighted by its NTF. In a way to keep only the major interests of a user, we consider only the top n terms according to their values. In our runs, we keep the top $n = 100$ terms in a profile.

Such profile is used as an expansion of initial query. In a way to reflect the relative importance of a term in a profile, we define a function $Exp(Q, u)$ that expands the query by a fixed number of terms, and the relative importance of each term in the profile is reflected as corresponding number of occurrences of this term in the expanded query. For instance, if the value corresponding to the term t in a user profile accounts for 40% of the occurrences of terms in the profile, and suppose that we fix the number of terms added in the query to 100, then the query will be expanded by $0.4 * 100 = 40$ occurrences of the term t. Then a BM25 retrieval is achieved on the documents corpus.

We noticed in SBS 2014 data that such expansion does not provides good results, but that the fusion of the results of such expanded queries and the usual content-only queries lead to better results. That is why we experimented such fusion with the parameters defined in table 2, where $NSProf$ denotes the parameters related to the non social profile fusion. The overall score for LIG_2 is:

$$\begin{aligned} RSV_{LIG\_2}(Q, d, u) = {} & \alpha_{BM25F}(Score_{BM25F}(Q, d) + b_{BM25F}) \\ & + \alpha_{LGD}(Score_{LGD}(Q, d) + b_{LGD}) \\ & + \alpha_{NSProf}(Score_{BM25}(Exp(Q, u), d) + b_{NSProf}) \end{aligned} \quad (2)$$

where α_{BM25F}, $\alpha_{LGD}, \alpha_{NSProf}$ are the relative importance of BM25 result list, LGD results list, the non social profil list respectively. Also, b_{BM25F}, b_{LGD}, b_{NSProf} are the respective bias of each results list. Moreover, the normalized scores use a horizon h of 1000, as presented in the table 2.

4.2 Friendship link generation

For the social user profile, we choose to generate "friendship" links between topic users and the non-topic users provided by SBS. To achieve that, we assume that what makes (topic or non-topic) users similar to others is their catalog and the ratings they provide. We represent then all the non-topics users as a text

Table 2. Parameters for the content + non social profile run **LIG_2**

	α	b	h
$BM25F$	0.4	0.5	1000
LGD	0.5	0.5	1000
$NSProf$	0.1	0.5	1000

document corresponding to concatenation of the document ids from the user catalog. We include the ratings (integer values) by using the ratings as the tf values for the number of occurrences of the documents ids.

To be able to find the non-topic users similar to topic users, we describe the users topics in the same way as the non-topic users as described above. Then we used the topic-users descriptions as queries on the corpus of non-topic users using a classical BM25 matching. For first experimentation, we filter the relationships to the top 2 most similar non-topic users for each topic user, and we plan to experiment the top k similar users in future works.

4.3 Usage of "friends"

Once 2 closer friends of a topic user are obtained, we apply a process similar to section 4 to generate the non-social profiles of the friends, and then we match the topic query with the friends profiles to get documents that match the query. The matching is computed as follows:

$$\begin{aligned} RSV_{LIG\_3}(Q,d,u) = {} & \alpha_{BM25F}(Score_{BM25F}(Q,d) + b_{BM25F}) \\ & + \alpha_{LGD}(Score_{LGD}(Q,d) + b_{LGD}) \\ & + \alpha_{NSProf}(Score_{Profil}(Q,d,u) + b_{NSProf}) \\ & + \alpha_{Fri1}(Score_{BM25}(Exp(Q,Fri1),d) + b_{Fri1}) \\ & + \alpha_{Fri2}(Score_{BM25}(Exp(Q,Fri2),d) + b_{Fri2}) \end{aligned} \quad (3)$$

The fusion parameters used for the officially submitted run **LIG_3** are given in table 3, with $Fri1$ and $Fri2$ the two friends of u.

Table 3. Parameters for the content + Non Social profile + Friends profiles run **LIG_3**

	α	b	h
$BM25F$	0.35	0.5	1000
LGD	0.45	0.5	1000
$NSProf$	0.1	0.6	1000
$Fri1$	0.05	0.6	1000
$Fri2$	0.05	0.5	1000

5 Documents given as "examples" (runs LIG_4, LIG_5 and LIG_6)

One important point to notice is that the post processing of the obtained results have a dramatic impact on the results. For instance, as the initial corpus ids (isbn) are not the ones on which the results are evaluated (LibraryThing ids), and because of potential duplicates generated, it is not obvious to handle the translation. Our approach for such duplicate removal was the same that is provided by the organizers of SBS.

Additionally, for our runs for SBS 2015, we focused on integrating the users examples to post-process the queries. Our idea was that a user might be interested if he finds as answers documents that he read and that he appreciated, as this would be an indicator that the system is providing relevant documents to him. We declined this hypothesis in two ways:

Reranking: Achieve a reranking where the documents a user likes are boosted, and the documents he dislikes are removed for the result. After a "Zero-one" score normalization [8] between 0 and 1 of the overall score, we add 1 for the documents that the user likes and set the score to 0 for the documents that he dislikes. This process is then a post-processing that is run after the fusion, and is the result of our official run **LIG_4**. It is worth noting that, if several retrieved documents are liked by the user, their relative initial ranking is preserved;

Relevance Feedback: Define a relevance feedback, positive for the documents that the topic user likes, and negative for the documents he does not like. We achieved such relevance feedback on the Log logistic run LGD for our official run **LIG_5**, and also on both content-runs, i.e., BM25F and Log logistic, for our official run **LIG_6**. The relevance feedback uses all the positive documents and selects the top 10 terms according to the default selection of Terrier [5].

6 Results

We present here two elements. First, we list the official results obtained for our 6 runs officially submitted to SBS 2015. Second, we discuss additional results generated after the release of the SBS 2015 qrels, presenting the impact of the "biases" we used (see section 2) over "unbiased" results.

6.1 Official Results

We comment here mainly the lines of table 4 corresponding to boldfaced run ids that use nor reranking neither relevance feedback. We notice then that the impact of the user non-social profile is clearly beneficial, however the p-value of a bilateral paired Student t-test on LIG_1 versus LIG_2 equal 5.45%, thus with

a significance threshold of 5% this difference is not statistically significant. Such value is even larger between LIG_1 and LIG_3. We notice a slight improvement of nDCG@10 results when integrating the 2 best "friends", however our generation or usage of relationships between users does not seem to be effective enough. According to what we defined for our fusion (see section 2), we also notice that the weights assigned to the friends are very small, 0.05. With higher relative values the results degrade. So we conclude for now that our proposal does not outperform the integration of non-social user information.

Table 4. Official results for the 6 **LIG** runs

Rank	Run	nDCG@10	MRR	MAP	R@1000	Profiles
6	**LIG_3**	0.098	0.189	0.069	0.514	yes
7	**LIG_2**	0.096	0.185	0.069	0.514	no
8	LIG_4	0.095	0.181	0.068	0.514	yes
13	LIG_5	0.093	0.179	0.067	0.515	yes
14	LIG_6	0.092	0.174	0.067	0.513	yes
15	**LIG_1**	0.090	0.173	0.063	0.508	no

As we see on table 4, the results with reranking or relevance feedback lower the quality of the results, but these elements are related to our interpretation of the catalogs and examples that are incompatible with the interpretation of the SBS organizers. In fact, our interpretation was somewhat the exact contrary of what decided the SBS organizers (they choose that the catalog + examples must not be part of the result), this explains why these additional runs behave worse than our initial runs.

6.2 Impact of biases

We describe in table 5 the impact of using the biases as defined in section 2. To be fair compared to the official results, we choose to only remove the bias from the configurations used for the official runs and to compare the relative gain or loss (between parentheses) with respect to the biased respective runs from table 4. In this table, the values use 3 digit precision numbers, where the percentages are computed on 4 digit precision numbers. We notice that the effect of the bias are positive for all the measures for the runs **LIG_2** and **LIG_3**, and have almost no effect of the content only **LIG_1** run. So, the effect of the biases seem to be more positive as we fuse many lists.

7 Conclusion

We presented in this paper the experiments that were conducted for the participation of LIG to the SBS 2015 lab evaluation. Our main finding, according to our integration of non-social and social profiles, is that the use of non-social

Table 5. Unofficial results without bias

Run	nDCG@10	MRR	MAP	R@1000
LIG_1 no bias	0.090 (0.0 %)	0.173 (-1,2 %)	0.063 (0.0 %)	0.507 (-0.1 %)
LIG_2 no bias	0.096 (-0.8 %)	0.182 (-1.6 %)	0.068 (-1.0 %)	0.513 (-0.2 %)
LIG_3 no bias	0.097 (-0.7 %)	0.181 (-4.2 %)	0.068 (-3.0 %)	0.508 (-1.2 %)

profile has a clear positive impact on the quality of the retrieval, where the integration of generated friendship relationships does not really increase the quality of the system provided. One important conclusion that we draw from the SBS experiments is that the post processing of results has a dramatic impact on the quality of the results, and then must be carefully studied.

The experiments reported here depict our first steps to grasp the complexity of personalized information retrieval in social context, and many efforts will focus on refining and characterizing the numerous elements involved in such retrieval process.

Acknowledgment

This work is supported by Région Rhône-Alpes through the ReSPIr project.

References

1. Cai, Y., Li, Q.: Personalized search by tag-based user profile and resource profile in collaborative tagging systems. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. pp. 969–978. CIKM 10, ACM, New York, NY, USA (2010)
2. Clinchant, S., Gaussier, E.: A Log-Logistic Model for Information Retrieval. In: 18th ACM Conference on Information and Knowledge Management. CIKM 10, vol. 14, pp. 5–25. Hong-Kong, China (2009)
3. Hafsi, M., Géry, M., Beigbeder, M.: LaHC at INEX 2014: Social book search track. In: Working Notes for CLEF 2014 Conference. pp. 514–520 (2014)
4. Koolen, M., Bogers, T., Kamps, J., Kazai, G., Preminger, M.: Overview of the INEX 2014 social book search track. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. pp. 462–479 (2014)
5. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. In: Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006) (2006)
6. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Found. Trends Inf. Retr. 3(4), 333–389 (2009)
7. Shaw, J.A., Fox, E.A.: Combination of multiple searches. In: The Second Text REtrieval Conference (TREC-2). pp. 243–252 (1994)
8. Wu, S., Crestani, F., Bi, Y.: Evaluating score normalization methods in data fusion. In: Ng, H., Leong, M.K., Kan, M.Y., Ji, D. (eds.) Information Retrieval Technology - Third Asia Information Retrieval Symposium, AIRS 2006. vol. 4182, pp. 642–648. Springer Berlin Heidelberg (2006)