Exploring combinations of sources for interaction features for document re-ranking

Conference or Workshop Item

Exploring Combinations of Sources for Interaction Features for Document Re-ranking

Emanuele Di Buccio
Department of Information Engineering
University of Padua, Italy
dibuccio@dei.unipd.it

Massimo Melucci
Department of Information Engineering
University of Padua, Italy
melo@dei.unipd.it

Dawei Song
School of Computing
The Robert Gordon University, UK
d.song@rgu.ac.uk

ABSTRACT

The behavior of the user when interacting with a result page or the corresponding landing documents is a possible source of evidence that Information Retrieval (IR) systems can exploit to assist the user when searching for information. Interaction features can be adopted as evidence to model the user behavior, thus making it usable to assist relevance prediction. One issue when dealing with interaction features is the selection of the sources from which these features are distilled. Individual users and group of users which perform a similar task or look for information matching the same query are possible sources. This paper will focus on these two sources, particularly investigating group of users searching for the same topic as source for interaction features to be used as an alternative to, or in combination with, individual users. The objective of this work is to investigate the impact of diverse combinations of these sources on the retrieval effectiveness, specifically when interaction features are used as evidence to support document re-ranking.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Relevance feedback, Search Process.

General Terms

Experimentation, Human Factors.

Keywords

Interaction Features, Implicit Relevance Feedback.

1. INTRODUCTION

A potential source of evidence to support IR systems when predicting relevance is the behavior of the user when interacting with result pages or the documents which results refer to. Implicit Feedback techniques [1] exploit features that can be gathered by monitoring the user behavior, e.g. interaction features, as indicators of the user interests or intents.

Interaction feature values can be distilled from diverse sources, and the selection of the source for features can affect the effectiveness of implicit feedback techniques. Individual users and group of interrelated users are possible sources from which interaction feature values can be observed. These sources have been investigated as alternative choices or used in combination because of their impact on the reliability and the availability of implicit features. For instance, in [2] the authors investigated display-time thresholds as implicit indicator of relevance. The obtained results showed that display-time, when considered in isolation and with regard to an individual user, may be hardly usable to support prediction; differently, it was a more consistent indicator when the threshold was learned from multiple subjects sharing a common task. In [3] user behavior models to support web search were learned by exploiting simultaneously feature values derived from the individual's behavior and those aggregated across all the users and search session for each query-URL pair, thus reducing the impact of individual variation in behavior. In [4] the authors investigated the impact of aggregating personalized scores per group formed according to diverse criteria on personalization algorithms when a small amount of personal data is available.

This work considers the scenario where a user interacts with some of the results obtained by a first search. The features gathered from this interaction can be exploited to obtain a representation of the information need which refines and complement the initial one, e.g. a textual query, as a new *dimension* of the user need representation. This dimension can be then adopted for document re-ranking. Beside the representation for the dimension, also documents need to be described in terms of the features observed from the user behavior. Two sources of evidence have been considered in the research work reported in this paper: individual users and groups of users. Since two source for features are considered, and a representation both for the dimension and the documents is required, that leads to different possible source combinations. For instance, the dimension can be modeled by the features gathered from the individuals, thus obtaining a personal user behavior model, and documents can be represented in terms of group data; this combination makes possible a representation in terms of interaction features for documents unseen by the individual users from which the evidence cannot be observed. Since other combinations are possible, this paper investigates the following research question: *Which is the best combination of source of features for modeling and exploiting the user behavior dimension for document re-ranking in terms of retrieval effectiveness?*

HCIR 2010, August 22, 2010, New Brunswick, NJ USA

2. METHODOLOGY

2.1 Methodology Description

The methodology adopted in this paper to exploit user interaction features for document re-ranking represents the user behavior dimension and the documents as vector subspaces according to the formalism proposed in [5]. The basic rationale is to map the collected data, prepared in a matrix, in a new vector space basis. The vector subspace spanned by this basis is the model of the dimension. The mapping is a matrix transformation technique which extracts information about our dimension from the collected data. For instance, if our hypothesis is that a dimension of the user information need can be represented by the correlation among interaction features, a technique like Principal Component Analysis (PCA) can be adopted. This is actually the approach proposed in [6] and adopted in this paper.

Also the documents to be re-ranked need to be described as subspaces; in this work they are represented as one dimensional subspaces, namely vectors, of the interaction feature values distilled from the users or the user groups behavior.

Once a representation in terms of subspaces has been built both for the dimension and the documents, the distance among the subspaces provides a measure of the degree to which the documents, represented with regard to the source, e.g. the user behavior, satisfy the dimension of the information need representation corresponding to that source.

Differently from [6] where user behavior models were used to support query expansion, we will focus on the impact of the diverse source combinations on document re-ranking.

2.2 Combinations of Sources for Features

The adopted methodology requires a representation both for the dimension of the information need and for the documents to re-rank with regard to the considered source. Since we are considering two distinct sources for features and two representations are required, this leads to four possible combinations of sources `X/Y`, where `X` denotes the source for the dimension and `Y` that for document representation – `X` or `Y` is either `P` (personal) or `G` (group). The `P/P` combination refers to the case where the features gathered from the individual user – i.e. its *personal data* – when searching for a specific topic are adopted both for modeling the dimension and for representing documents. The `P/G` combination refers to the case where personal data are adopted for modeling the dimension, while the data gathered from a group of users searching for the considered topic are adopted for documents representation. The remaining two combinations have analogous meaning.

The experiments reported in the next section aim at investigating the impact of the above source combinations on the retrieval effectiveness, specifically when they are adopted to support user behavior-based document re-ranking.

3. EXPERIMENTS

3.1 Evaluation Methodology

The basic rationale underlying the methodology to investigate the above research question was to observe the behavior of the user when visiting the first n results, and then use this evidence to model it, specifically as a vector subspace; this subspace representation was then used for re-ranking the top m documents provided by the baseline.

In particular, the evaluation methodology consisted of the following steps, that were performed for each topic-user pair:

1. Selection of the combination of the source for features, that is `P/P`, `P/G`, `G/P`, or `G/G`.
2. Collection of the features from the first $n = 3$ visited documents[1]. The collected features are prepared in a matrix $F \in \mathbb{R}^{n \times k}$, where k is the number of features collected from the n visited documents.
3. Modeling the dimension of the information need representation by extracting possible behavioral patterns by applying PCA on F. The result of the application of this technique is an orthonormal basis – one basis vector $\mathbf{b}$ for each pattern. Patterns, namely eigenvectors, associated to non-null eigenvalues are tested one at a time as possible models for the dimension – the model of the dimension is the subspace $L(\{\mathbf{b}\})$ spanned by $\{\mathbf{b}\}$, namely a one-dimensional subspace.
4. Representation of the documents in terms of features gathered from the source selected at step 1. Each document is represented as a vector $\mathbf{y}$ of k features.
5. Re-ranking of the top $m = 10$ results of the baseline list according to the measure $m_{\mathbf{b}}(\mathbf{y}) = \mathbf{y}^T \cdot \mathbf{P}_{L(\{\mathbf{b}\})} \cdot \mathbf{y}$, where $\mathbf{P}_{L(\{\mathbf{b}\})} = \mathbf{b} \cdot \mathbf{b}^T$ is the projector onto $L(\{\mathbf{b}\})$. $m_{\mathbf{b}}(\mathbf{y})$ provides a measure of the degree to which the document representation satisfies the dimension model.
6. Computation of the NDCG@10 for the new result list obtained after document re-ranking using the gains provided by the user for the considered topic[2].

When the *group* was adopted as source for features for modeling the dimension, namely in the `G/`$\cdot$ or $\cdot$`/G` combinations, the value $f^G_{i,u',d,t}$ of a feature i for a specific user-topic-document (u', t, d) triple was computed as

$$f^G_{i,u',d,t} = \frac{1}{|G| - 1} \sum_{u \in G \text{ and } u \neq u'} f^I_{i,u,d,t}$$

where G denotes the group constituted by all the users which visited the document d with regard to the topic t and $f^I_{i,u,d,t}$ the feature value observed for a specific individual u with regard to (d, t). In other words the group value of a feature for a specific user u' was obtained as the average value of the feature values observed for the other users in G. The reason for this choice was to test if the evidence gathered from users in G other than u' can "substitute" the feature values for the document unseen by the user.

3.2 Dataset

Addressing the considered research question requires a dataset constituted by a set of topics, the properties of the results and the documents to re-rank for each topic, the features when interacting with them, and finally explicit judgments of the users for each topic-document pair.

[1]The order in which the users visited the documents in the study described in Section 3.2 was not necessarily the displayed order, so the visited order is adopted.

[2]DCG is computed according to the alternative formulation reported in [7], namely $\sum_i (2^{r(i)} - 1)/\log(i + 1)$, where $r(i)$ is the relevance of the document at position i. The normalization factor is the DCG of the perfect ranking.

Difficulty	# of Relevant Docs	Topics
High	1/2	506 - 517 - 518 - 543 - 546
Medium	3/4/5	501 - 502 - 504 - 536 - 550
Low	6/7/8/9/10	509 - 510 - 511 - 544 - 549

Table 1: Topic bins.

Feature	Description
Features observed from document/browser window	
query terms in title	number of topic terms displayed in the title of the corresponding result
ddepth	depth of the browser window when examining the document
dwidth	width of the browser window when examining the document
doc-length	length of the document (number of terms)
Feature observed from the user behavior	
display-time	time the user spent on the page in its first visit
scroll-down	number of actions to scroll down the document performed both by page down and mouse scroll
scroll-up	number of actions to scroll up the document performed both by page up and mouse scroll
sdepth	maximum depth of the page achieved by scrolling down, starting from ddepth

Table 2: Features adopted to model the *user behavior* dimension and to represent documents.

A dataset with this information has been gathered through a user study which involved fifteen people which were asked to assess the top 10 retrieved results in response to nine several assigned topics and to assess their relevance with a four-graded scale. We adopted the WT10g test collection and the ad-hoc topics of the TREC 2001 Web Track. The collection was indexed by the Lemur Toolkit[3]; english stopwords were removed and the Porter stemmer was adopted. Kullback-Leibler (KL) Divergence was adopted to rank documents because of its effectiveness in the TREC 2001 Web Track [8]. Then the top 10 documents were considered for each topic. A subset of the fifty topics were adopted in the study: topics were divided in three bins according to their difficulty – see Table 1 – where the measure of difficulty was the number of relevant documents in the top 10 – here relevance refers to the judgments provided by TREC assessors. We randomly selected five topics per bin, thus obtaining fifteen distinct topics; then three distinct groups of nine queries were built and distributed among the users; each user was asked to assess topics in one group.

The assessment was performed by a web application which displayed for each topic its description, the list of the titles of top 10 results for that topic, and when a result title was clicked by the user, the content of the document corresponding to that title. Both client-side and server-side functionalities were adopted to gather features, specifically those reported in Table 2 – these features are those adopted to prepare the matrix F in step 2 of Section 3.1.

To gather explicit judgments beside each title a drop down menu was available to select the relevance degree of the document corresponding to that title – these judgments are those used in step 6 of Section 3.1. Some users did not assess all the documents in the result list for some topics. For this reason, in regard to the objective of this paper, only the user behavior of thirteen among the fifteen users were considered in this work, for a total of 79 (user,topic) pairs and 790 entries where each entry refers to the visit of a specific user to a particular document with regard to a topic.

[3]`http://www.lemurproject.org/lemur/`

4. RESULTS

Table 3 reports the average and the median NDCG@10 computed over all the entries for the different combinations. There is no significant difference among the contributions of the diverse combinations. This is confirmed by the Wilcoxon signed rank test performed – with a 95% confidence interval – between the NDCG values obtained using the different source combinations. The Wilcoxon signed rank test was adopted since the Shapiro-Wilk test showed that normality cannot be assumed for the NDCG's values obtained by the different combinations.

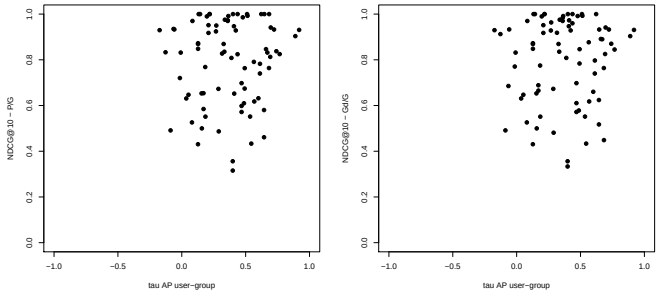
The `G/-` combinations (`G/P` and `G/G`) performed worse than the correspondent `P/-` combinations (`P/P` and `P/G`). A possible reason for the low performance of `G/-` was the adoption of the average values of the features over the group to model the dimension. In order to investigate this hypothesis we considered another combination labeled as `Gd/G`. In this combination, as for the other `-/G` cases, the evidence adopted to represent the documents with regard to a topic is obtained by computing the average values of the features over all the users other than the user under consideration that assessed that topic. The `Gd` label denotes that the model was obtained by applying PCA to a document-by-feature matrix where the documents of the diverse users were considered as distinct evidence. For instance, if the system was supporting user1 when searching for topic 502, the feature matrix adopted as evidence was $F \in \mathbb{R}^{(n \cdot 5) \times k}$, where k is the number of features, n is the number of visited documents, and 5 is the number of users other than user1 that searched for topic 502 in the collected dataset.

The average and the median NDCG@10 values for the `Gd/G` combination computed over all the users and all the topics was higher than that obtained for the other combinations. The Wilcoxon signed rank test showed that the improvement `Gd/G` respect to `P/G`, `G/G` and `G/P` was significant with a 95% confidence interval – ($V = 399$, p-value $= 0.002$) for `P/G`, ($V = 1746.5$, p-value $= 0.03$) for `G/P`, and ($V = 560$, p-value $= 0.00001$) for `G/G`. This result confirms the negative impact of using average feature values to prepare the matrix F. The best performance achieved by the `Gd/G` case may be also due to the larger amount of data adopted for modeling the dimension: the PCA-based approach seems to benefit from a larger number of observations. Moreover, the better results obtained for the `P/G` and the `Gd/G` cases suggest that group behavior features can substitute personal features for document representation, thus making personalized IRF feasible despite the data sparsity observed when the interaction features are collected on a per-user basis.

The results observed for the `P/G` and the `Gd/G` could be due to the level of agreement between what the user and the group perceived as relevant. In order to investigate this hypothesis, the NDCG@10's for these combinations were plotted against the τ_{AP} [9] computed among the ideal individual ranking and the ideal group ranking for each (user,topic) pair. The ideal ranking for a user was obtained by ranking documents by the gain he provided. For the group ideal ranking, the gain of each document was obtained as the sum of the gains provided for that document by the users in the group, as proposed in [10] to compute NDCG for a group of

	Baseline	Source combinations				
NDCG@10	KL	P/P	P/G	G/P	G/G	Gd/G
Average	0.765	0.765	0.791	0.759	0.777	0.797
Median	0.838	0.817	0.832	0.799	0.825	0.869

Table 3: Average and median NDCG@10 computed over all the (user,topic) pairs.



(a) P/G (b) Gd/G

Figure 1: NDCG@10 for the P/G (Fig. 1a) and the Gd/G (Fig. 1b) combination compared with τ_{AP} between user and group (not including the user) gains.

	Baseline	Source combinations		Increment (%)	
User	KL	P/G	Gd/G	Δ_{PG-KL}	Δ_{GdG-KL}
user1	0.760	0.766	0.810	0.784	6.520
user2	0.688	0.844	0.885	22.662	28.744
user3	0.726	0.729	0.747	0.445	2.949
user5	0.798	0.798	0.803	0.054	0.671
user7	0.775	0.823	0.699	6.283	-9.700
user8	0.737	0.758	0.766	2.853	3.942
user9	0.792	0.759	0.770	-4.134	-2.724
user10	0.850	0.886	0.847	4.314	-0.341
user11	0.799	0.776	0.776	-2.825	-2.825
user12	0.866	0.756	0.767	-12.681	-11.381
user13	0.676	0.839	0.849	24.103	25.552
user15	0.839	0.820	0.849	-2.302	1.102
user16	0.670	0.733	0.745	9.284	11.147

Table 4: NDCG@10 per user

users. The scatter-plots depicted in Figure 1a and Figure 1b, which refer respectively to the P/G and the Gd/G case, suggest that the agreement did not impact of the performance of the two combinations.

When compared with the baseline, the diverse combinations did not provide on average a significant improvement respect to the baseline KL – the best performing combination Gd/G provided an improvement of 4.18% and 3.70% (V = 1698, p-value = 0.07) respectively in terms of average and median NDCG@10. Table 4 reports the average NDCG@10 value for each user with regard to the baseline and the two best performing combinations. The improvement is not consistent among the users, and for some of them, e.g. user12, the user behavior-based re-ranking negatively affected the initial ranking. These results suggests that further research work is needed to understand why and when these features are an usable source for improving retrieval effectiveness.

5. CONCLUDING REMARKS

This paper has investigated the impact of the selection of the source for interaction features on document re-ranking, when those features are used to obtain a usable representation of the information need and of the documents.

The results of the experiments carried out in this work showed that the contribution of the diverse combinations is comparable, although the combinations where group data were adopted for document representation performed slightly better. In particular, significant difference with the other combinations was observed only for the combination where the model was learned from feature values of the individual constituting the group considered as individual entries. These results suggest that group data can be a good source for document representation, thus making possible a representation also for documents unseen by the individual users.

Since groups were constituted by users assessing the same topic, we investigated if the comparable results obtained for individual and group-based representations were due to the agreement between individual's and group gains, but no relationship between NDCG's and agreement was found.

Future investigation will be focused on more realistic grouping criteria than considering users with the same information need – the entire topic description was shown to the users.

The strategy adopted in this paper to extract behavioral patterns requires the manual selection of the best performing pattern. Although this approach was appropriate for exploring source combinations, different techniques are needed and will be investigated to automatically support individual's.

6. REFERENCES

[1] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, 2003.

[2] R. W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of CIKM '06*, pages 297–306, New York, NY, USA, 2006. ACM.

[3] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proceedings of SIGIR '06*, pages 3–10, New York, NY, USA, 2006. ACM.

[4] J. Teevan, M. Ringel Morris, and S. Bush. Discovering and using groups to improve personalized search. In *Proceedings of WSDM '09*, Barcelona, Spain, 2009.

[5] M. Melucci. A basis for information retrieval in context. *ACM TOIS*, 26(3):1–41, 2008.

[6] M. Melucci and R. W. White. Utilizing a geometry of context for enhanced implicit feedback. In *Proceedings of CIKM 2007*, Lisbon, Portugal, 2007.

[7] W. B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison Wesley, 2009.

[8] P. Ogilvie and J. Callan. Experiments using the Lemur toolkit. In *Proceedings of TREC-10*, Gaithersburg, MD, USA, 2001.

[9] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of SIGIR '08*, pages 587–594, New York, NY, USA, 2008. ACM.

[10] J. Teevan, S. T. Dumais, and E. Horvitz. Potential for personalization. *ACM TOCHI*, 17(1):1–31, 2010.