# Exploiting individual users and user groups interaction features: methodology and infrastructure design

Emanuele Di Buccio and Massimo Melucci

Department of Information Engineering, University of Padua, Italy
{dibuccio,melo}@dei.unipd.it

**Abstract.** The user may be a source of evidence for supporting information access through Digital Library (DL) systems. In particular, the features gathered while monitoring the interaction between the user and a DL system can be used as implicit indicators of the user interests. However, each user has his own style of interaction and a feature which is a reliable indicator with regard to one user may be no longer reliable when referred to another user. This suggests the need to develop personalized approaches for each user which are tailored for each search task. Nevertheless, the behavior of a group of interrelated users, e.g. performing the same task, may improve the contribution provided by the personal behavior; for instance, some interaction features, if considered individually, are more reliable with regard to a group of users. This paper introduces a methodology for exploiting both the behavior of individual users and group of users as sources of evidence. The paper also introduces a software infrastructure implementing the methodology. The methodology is mainly based on a geometric framework while the software infrastructure is based on a partially decentralized Peer-To-Peer (P2P) network, thus permitting the management of different sources of evidence.

## 1 Introduction

The heterogeneity and the size of the data made available by Digital Library (DL) systems, the high number of users and user groups accessing such systems, and the variety of the information needs require strategies which help the users to access relevant information. This paper addresses the design of strategies for information access both at methodological and infrastructural level. First, the paper introduces a methodology for exploiting a variety of sources of evidence for enhancing information access. The specific sources considered in this paper are the features gathered by monitoring the behavior of a user or a group of users somehow interrelated; Section 2 provides motivations for this choice. The methodology is structured in four steps, the aim of which is to support the selection of the sources, to collect the evidence, to model the sources and their relationships in order to make them usable, and to assist their use for accessing relevant information. Sources and relationships are modeled and then exploited through a previously proposed geometric framework, which provides a uniform and usable representation for them in terms of vector space subspaces. A detailed description of the methodology will be provided in Section 4 and Section 5. This paper also provides a contribution at the infrastructural level as the methodology is complemented by the design of a partially decentralized infrastructure based on the P2P paradigm which allows data about the behavior of the individual users and users groups to be managed. Section 3 provides motivations for the infrastructure. Section 5 explains its role in regard to the proposed methodology.

## 2 User behavior: a source to support information access

Information about user behavior when interacting with DL systems can be a useful source of evidence to improve the systems capability of providing the relevant documents which the user is looking for. The user can explicitly be asked to provide information about his intents or interests. For instance, the user can explicitly rate images, audio files, videos, or label documents with tags to describe them. This information is a valuable source for building user profiles which can be adopted to suggest other possible results of interest. Explicit feedback has been shown to be an effective strategy to support the user during searching, e.g. to refine or expand queries.

Despite its effectiveness, explicit feedback is not always perceived useful by users due to the effort required by direct involvement of the user. An alternative is the adoption of implicit indicators as a source of evidence for relevance feedback. The expression "implicit indicators" refers to those features collected without direct involvement by the user; examples include user interaction features, e.g. display-time, click-through data or the features gathered during the user study described in [1]. Implicit Relevance Feedback (IRF) algorithms [2] exploit implicit features to provide feedback to the user, thus preserving the benefits of explicit relevance feedback and removing its burdens.

When user behavior is used as a source of evidence for IRF algorithms, the behavior features can be collected with regard to the individual user or group of interrelated users — e.g. a group of users who perform the same task, submit the same topic or belong to a social network. When the individual user is adopted as a source of evidence, the features have *personal granularity*[1], e.g. the display-times or the click-through data are gathered only by the observation of the considered user. In the remainder of this paper we will refer to the set of personal granularity features, or to the result of their arrangement or combination, as the *personal behavior dimension*. The approach proposed in [3], for instance, simultaneously exploits multiple interaction features with personal granularity for developing enhanced implicit feedback models personalized for each user and tailored for each search task. One of the issues when exploiting personal information is privacy concerns – if a centralized system is adopted, the user may be reluctant to use personalization if it requires providing personal data.

As mentioned above, not only the individual user, but also groups of users explicitly or implicitly formed can be used as a source of evidence for IRF algo-

[1] The *granularity* of the features is the level of detail at which the features are observed.

rithms – similar to the personal features, the set of behavior features referring to the group can be named *group behavior dimension.* One of the reasons for exploiting a group of users as a source is, for instance, that some interaction features, e.g. display-time, are not reliable indicators when considered individually and with regard to a single user performing a specific task; whereas multiple users performing the same task tend to interact more consistently [4]. Implicit features at *group granularity* were adopted in [5], where the *hit-matrix* which stores the search behavior data – the element $(i, j)$ of the matrix represents the number of times the page $j$ was accessed with regard to the query $i$ – maintains information at the *community* granularity level – in [5] a "community" represents a set of users working in the same company or accessing an interest-specific portal. Another reason for exploiting group evidence is that the contribution provided by the group can be different from that provided by the individual user – e.g. in [6] the information gathered by groups of users is adopted to enhance personalization.

## 3 An infrastructure to manage user behavioral features

The collection and the aggregation of user information for enhancing information access may cause concerns about privacy preservation. In [7] the authors proposed several levels of privacy protection in personalized search and then examined several software architecture for personalization with regard to the defined privacy levels. The rationale of the infrastructure proposed in this paper is close to the architecture named *client-server cooperative personalization* in [7]. In this architecture the user information is stored on the client side; at query time the client extracts the information from the user profiles and sends it to the server, namely the search engine, where personalization is performed.

The specific infrastructure adopted in this paper is based on the SPINA software architecture [8]. Unlike other papers, a partially decentralized infrastructure based on the P2P paradigm is proposed in this paper. SPINA was designed and developed for providing indexing and retrieval of unstructured documents distributed across P2P networks which are unstructured (i.e. no DHT-like data structures), hybrid (i.e. the simultaneous presence of peers and ultra-peers) and hierarchical (i.e. each peer refers to one and only one ultra-peer which serves a group of peers acting as a hub/router for queries sent by a peer in its group). Ultra-peers are peers with previously established attributes – for example with more CPU, bandwidth or disk than the others – which are *dynamically* elected from normal peers. An instance of this kind of network is depicted in Figure 1. The retrieval strategy adopted in SPINA exploits indexes at different granularities, particularly with peer and ultra-peer granularity. In these indexes the elements in the posting list associated with each feature are the identifier of the peer (or the ultra-peer) and the weight of the feature in that peer (or ultra-peer). These weights are computed by the aggregation of the statistics of the features according to the level hierarchy.

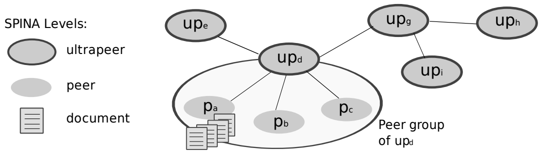

Fig. 1: SPINA Levels.

The functionalities provided by SPINA to manage the features at the two granularity levels can support the methodology proposed in this paper based on the two different behavioral feature granularities, i.e. individual user and group granularity. In particular, the rationale is that features with personal behavior granularity are stored in the peers, while the group behavior granularity features are stored on the ultra-peer side and are obtained as aggregation of the features stored in the peers which refer to that ultra-peer. The ultra-peer acts as a proxy server, thus all the users interacting with peers in the ultra-peer group are interpreted as an individual user.

## 4 A methodology to exploit user behavior

Information access and retrieval systems can exploit the evidence provided by different sources for improving retrieval effectiveness. Although mostly used metadata, content and structure of the documents are only a subset of the available sources. Instances of other sources of evidence are the search task, the specific topic within the task [1], or the location. The complexity of the design and the development of approaches to exploit different sources is not only due to the number of sources involved, but also to the relationships existing among such sources. Let us consider a user who is looking for restaurants in London and interacts with the results returned by the DL system. If the keyword "jazz" appears in the first result selected by the user – e.g. in the snippet, the title or the url related to the result – the user behavior probably suggests he is more interested in jazz restaurants than in generic ones. Therefore, a relationship can exist between the two sources considered, that is, the user behavior and the terms appearing near the query features in the displayed results.

The point here is that the different sources are not necessarily independent of each other – the features observed from a source (e.g. the behavior) are "entangled" with the features observed from another source (e.g. the particular meaning of a query feature in the selected results). The design of different approaches, one for each source, may fail to deal with the challenge of modeling the relationships existing among sources and consequently fail to exploit them. The methodology introduced in [9] and refined in this work, aims at addressing this problem through a uniform and usable representation of the different sources and their relationships. The methodology is structured in four steps which are depicted in Figure 2 with regard to a specific source, namely user behavior.

The first step of the methodology is the *selection of the sources* adopted to support information access. This choice is not unique, but can depend on the

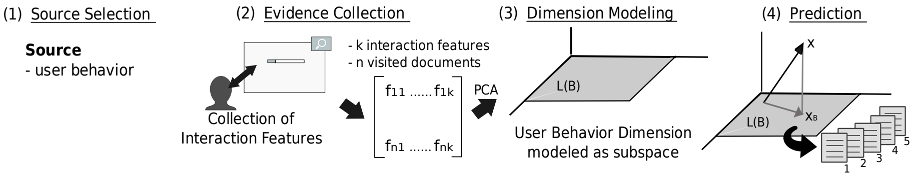

Fig. 2: Methodology Steps for User Behavior Dimension.

specific domain of application: the sources involved when the tool is developed for a mobile device are different from those which can be involved when designing a desktop search tool or when addressing the problem of enterprise search. The specific source considered in this work is the behavior of the user when interacting with the results returned by the system. In accordance with the terminology adopted in Section 2, this source will be named as *user behavior dimension*.

Once the sources have been selected, the dimensions identified to describe such sources have to be modeled. The model of a dimension is built around the evidence collected by monitoring the sources related to this dimension. Feature selection and gathering constitutes a preliminary step to support dimensions modeling: this step will be named *evidence collection*. The selection of the suitable features is part of the design to implement the methodology, since it affects the modeling step. Indeed, a feature or a set of features have to be a reliable indicator of the user needs, intents or interest. Interaction features have been selected in this work to describe the user behavior dimension. Moreover, the feature selection step provides some requirements to the design of the tool which practically will gather the features. For instance, interaction features can be monitored by an extension of a browser or a plug-in for a media-player application.

The collected features constitute a representation of the source, but this representation is not always directly usable: features often appear as rough and noisy data. The purpose of the *dimension modeling* step is to model a source in order to make it usable, thus supporting access to relevant information. In this work the modeling step is addressed by the adoption of the geometric framework proposed in [10], whose rationale is to use vector space subspace as a construct to model a dimension or a set of dimensions. In order to understand the rationale of this step, let us consider the result of the previous step when applied to the user behavior dimension. Let us suppose the user interaction behavior was monitored when accessing and interacting with the first $k$ results visited. The gathered interaction features can be prepared in a matrix $A$, whose $(i, j)$ element is the feature $j$ observed during the visit of document $i$ — see Step (2) in Figure 2. The matrix A is a vector-based representation of the observed data, but this representation does not always reveal the logical structure underlying the data. A matrix transformation technique can be adopted to reveal the logical structure: for instance in [3] Principal Component Analysis (PCA) of $A^T A$ is adopted to compute the vector space basis. A subset of the eigenvectors computed by PCA can be used as a basis to model the user behavior dimension: the basis spans a subspace which is the model of the dimension – e.g. the subspace $L(B)$ in Fig. 2.

The last step is the *prediction* of the documents which can be relevant to the user information need. This step can be performed by the adoption of the trace-based function adopted in [10], which provides a measure of the degree to which the modeled dimensions occur in a document. Documents can be ranked according to this measure, which has proven to be a probability measure. The rationale is to compute the distance between the vector representation of the documents in terms of features of the selected dimension – e.g. vector $\mathbf{x}$ depicted in step (4) of Figure 2 – and the subspace modeling the dimension(s) spanned by the computed vector space basis – e.g. the subspace $L(B)$ in Figure 2.

## 5 Implementing user behavior granularities

The methodology introduced in Section 4 is general and in order to apply it to a specific dimension an implementation of the four steps is required. This section aims at providing a possible implementation with regard to the user behavior as a source, moreover focusing on the infrastructure to support this implementation.

Let us consider a user who interacts with a DL node, namely a peer, in order to submit a query. Since in this work interaction features are adopted to model the user behavior dimension, the peer needs a monitoring tool able to gather information about the interaction between the user and the application to access information – e.g. if the application is developed to run in a web browser, this monitoring tool may be a browser extension. When the system returns the results to the user, the monitoring tool gathers interaction features with regard to the first $k$ visited documents and then this evidence is used to obtain a model of the user behavior dimension — a possible approach is the one proposed in [3] and mentioned in the description of the *dimension modeling* step, i.e. modeling a dimension as a vector space subspace. Once the subspace has been computed, the documents can be re-ranked by the trace-based function proposed in [10]. The documents to be ranked can be part of the local collection or distributed over Internet or across a P2P network. If only the local collection is considered, the user behavior can be useful only for the documents already seen: no evidence is available for the other documents, but the user profile can be enriched across different searches.

The methodology together with the infrastructure can provide a different contribution when documents other than those stored locally are accessed. SPINA provides not only functionalities to perform local search, but also to search documents distributed across a P2P network. In particular, a SPINA peer forwards the query to its referring ultra-peer, which according to the query features — e.g. keywords or audio patterns — selects and forwards the query to the most promising peers in its group — details about the algorithm can be found in [8]. The contacted peers locally retrieve the most promising documents, which are returned to the ultra-peer. Moreover, the ultra-peer can forward the query to its most promising neighboring ultra-peers which will iterate the search process in their groups and return the results to the starting ultra-peer.

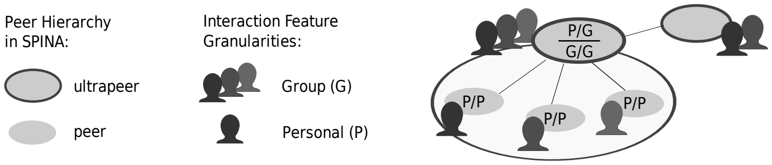

Fig. 3: SPINA to support the user behavior dimension.

Accessing the documents distributed across the network not only augments the contents available, but also improves the support provided by the user behavior dimension. Indeed, the retrieved documents may be accessed by other users, for instance when submitting similar queries or performing similar tasks. Each user can set its peer to periodically provide his behavioral features to its referring ultra-peer. The ultra-peer aggregates the contribution provided by such interaction features with the information already present with regard to the same task, similar topics, or in general according to the selected aggregation policy. For instance, the number of clicks on a specific document with regard to a particular user when performing a certain task can be added to the number of clicks stored in the ultra-peer with regard to the same document and the same task. This information will not refer to the behavior of the individual user, but to the behavior of the group of users interacting with peers in the ultra-peer group. Since the features at user granularity are not maintained in the ultra-peers, the user may be encouraged to provide its contribution in terms of interaction features.

These group granularity features can be used to complement the personal granularity features when the latter are not available for some documents which can be described in terms of group granularity interaction features. In particular, after $k$ documents have been visited, the peer sends the monitored interaction features with personal granularity to the ultra-peer, which builds a representation of the personal behavior dimension and uses the group granularity interaction features available to re-rank the documents against the personal interaction dimension. Finally, the ultra-peer returns the results to the requesting peer and updates the group granularity features with the features previously sent by the peer. This approach can be interpreted as a $P/G$ combination exploiting personal features ($P$) to model the dimension and the group granularity features ($G$) to represent the documents. By using the same labeling scheme, the feedback completely performed locally, namely in the peer, can be labeled as $P/P$ which exploits the personal features both to build the dimension model and to represent the documents. Another possible approach is the $G/G$ case which exploits the group granularity features both for dimension modeling and document representation – this approach is exploited at the ultra-peer level. Since each user has his own style of interaction, this combination based on the group behavior can be useful to diversify result suggestions. Figure 3 depicts the level of the peers hierarchy where each of the three different approaches – i.e. $P/P$, $P/G$ and $G/G$ – are performed.

## 6 Concluding Remarks

In this work we have proposed a methodology and an infrastructure which aims at supporting information access by exploiting user behavior described in terms of interaction features at different granularities, in particular, referred to individual users ($P$) and user groups ($G$). Some preliminary experiments conducted on the effectiveness of the different combinations of interaction sources, showed how the $P/P$ combination is the most effective one, followed by $P/G$ and $G/G$; we are currently investigating if the different combinations provide diverse contributions and how to combine them.

The main issue under investigation is the grouping criterion, since it affects the policy of aggregation of the personal interaction features in order to obtain the features with group granularity. A possible solution is grouping users by tasks [4, 6]; in our preliminary experiments the users were grouped according to the search task, but groups were manually defined. The issue to be addressed is how to automatically create groups; a solution to this issue may also suggest policies according to which the peers join peer groups, namely select ultra-peers. In regard to the current implementation of SPINA, behavioral features can be a source of evidence for supporting the result merging procedure. A result merging technique is required because the scores assigned to the documents according to their contents by the different peers are not directly comparable. Behavioral features can be adopted to re-rank the results in the ultra-peers and to present a final ranked list of results to the user.

## References

1. D. Kelly. Measuring online information seeking context, part 1: Background and method. *JASIST*, 57(13):1729–1739, 2006.
2. D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, 2003.
3. M. Melucci and R.W. White. Utilizing a geometry of context for enhanced implicit feedback. In *Proceedings of CIKM'07*, pages 273–282, 2007.
4. 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.
5. B. Smyth. A community-based approach to personalizing web search. *Computer*, 40(8):42–50, 2007.
6. J. Teevan, M. Ringel Morris, and S. Bush. Discovering and using groups to improve personalized search. In *Proceedings of WSDM'09*, Barcelona, Spain, 2009.
7. X. Shen, B. Tan, and C. Zhai. Privacy protection in personalized search. *SIGIR Forum*, 41(1):4–17, 2007.
8. E. Di Buccio, N. Ferro, and M. Melucci. Content-based Information Retrieval in SPINA. In *Proceedings of IRCDL2009*, Padova, Italy, 2008.
9. E. Di Buccio and M. Melucci. Towards a Methodology for Contextual Information Retrieval. In *Proceedings of the Workshop on Contextual Information Access, Seeking and Retrieval Evaluation*, Tolouse, France, 2009.
10. M. Melucci. A basis for information retrieval in context. *ACM Transaction on Information Systems*, 26(3):1–41, 2008.