Issues in Personalizing Information Retrieval

Gabriella Pasi

***Abstract*—This paper shortly discusses the main issues related to the problem of personalizing search. To overcome the "one size fits all" behavior of most search engines and Information Retrieval Systems, in recent years a great deal of research has addressed the problem of defining techniques aimed at tailoring the search outcome to the user context. This paper outlines the main issues related to the two basic problems beyond these approaches: context representation and definition of processes which exploit the context knowledge to improve the quality of the search outcome. Moreover some other important and related issues are mentioned, such as privacy, and evaluation.**

***Index Terms* — Information Retrieval, Personalization, Context Modeling, User Modeling.**

I. Introduction

IN recent years there has been an increasing research interest in the problem of contextualizing search to the aim of overcoming the limitations of the "one size fits all" paradigm, which is generally applied by Search Engines and Information Retrieval Systems (IRSs). By this paradigm the keyword-based query is considered as the only carrier of the users' information needs. As a consequence, the relevance estimate is system-centered, as the user context is not taken into account. Instead, a contextual Search Engine or IRS relies on a user-centered approach since it involves processes, techniques and algorithms that exploit as much contextual factors as possible in order to tailor the search results to users [6,14,19,27,28,37].

As it will be shown in section II, the key notion of context may have multiple interpretations in Information Retrieval (IR). It may be related to the characteristics and preferences of a specific user or group of users (in this case contextualization can be referred to as personalization), or it may be related to user geographic localization (when for example using a search engine on a smart-phone), or it may refer to the information that qualifies the content of a given document/web page (for example its author, its creation date, its format etc.).
The development and increasing use of tools that either help users to express their topical preferences, or automatically learn them, and the availability of devices and technologies that can detect both users' location (such as GPSs) and monitor users' actions, allow to capture the user's context, related to the considered interpretation or application in the attempt to contextualize search.

To the aim of modeling contextualized IR applications, a significant amount of research has addressed two main problems: how to model the user's context, and how to exploit it in the retrieval process in order to provide context-aware results. Several research works have offered possible solutions to the above problems, related to the considered interpretation of context, giving birth to some specific IR branches such as personalized IR, mobile IR, social IR. Although the specific techniques related to these branches vary (due to the nature of context that needs to be modelled), the common issue of context-based IR is to improve the quality of search by proposing to the user results tailored to the considered context.

In this paper a synthetic overview of some main issues in designing personalized approaches to Information Retrieval is presented. In section II the shift from the system centered approach to the user and context centered approach in IR is discussed. Section III aims at reporting on the issue of defining a formal user model; in section IV the approaches proposed in the literature to exploit the user context in search are classified and shortly described. Finally in section V the important issues of privacy and personalized systems evaluation are discussed.

II. From the System Centered Approach to a User Centered Approach to IR

Most Information Retrieval Systems and Search Engines rely on the so called system-centered approach, where the IRS behaves as a black box, which produces the same answer to the same query, independently on the user context. The notion of context in IR is well described in [36], and it may have several interpretations, ranging from user context (the central notion in context-based IR), to document context, spatio-temporal context, social context, etc. The identification of a specific context allows to identify information that can be usefully exploited to the aim of improving search effectiveness. For example, by user context we generally refer to the information characterizing a person (personal information) and his/her preferences. The personal information may include demographic and professional data; preferences of a person may range from topical preferences, taste preferences, etc. The spatio-temporal context is identified by information such as location, geographic coordinates etc.

If properly acquired, organized, and stored, the context-related information may be used to leverage the process aimed at identifying information relevant to a user need, beyond the mere usage of the user's query. To this aim a context model must be defined by a formal language, which is used to represent the information related to the context.

Gabriella Pasi is with the Università degli Studi di Milano Bicocca, Department of Informatics, Systems and Communication, Milano, Italy (phone: +39-02-64487847; e-mail: pasi@ disco.unimib.it).

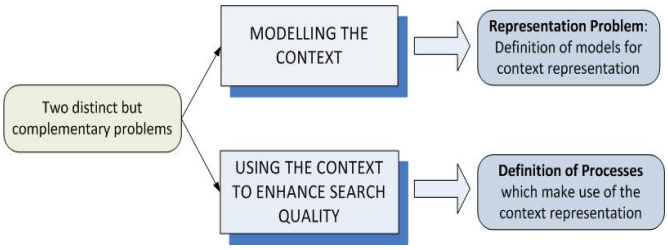

Fig. 1. The main processes involved in personalized IR.

Context-centered IR is an expression which can be used to encompass all tools, techniques and algorithms finalized at producing a search outcome (in response to a user's query), which is tailored to the specific context. This way the "one size fits all" approach is no more valid. When context is referred to the user context, we may talk about personalized IR.

The previous short introduction to the notion of context and its possible use in IR makes it evident that in order to implement a context dependent IR strategy, two main activities must be undertaken, as sketched in Fig.1. The prerequisite activity is of type knowledge representation, and is aimed at the definition of the context model. Such an activity comprises sub-activities such as the identification of the basic knowledge which characterizes the context, the choice of a formal language by which to represent this knowledge, and a strategy to update this knowledge (to adapt the representation to context variations). The second activity is aimed at defining processes (algorithms), which, based on both the knowledge represented in the context representation and the user query, are finalized to produce as a search outcome an estimate of document relevance which takes into account the context dimension(s). In other words, the context is used to leverage the effectiveness of the search outcome. As it will be explained in section III this can be done by different approaches, which can be classified depending on the way in which the contextual information is exploited.

While in this section we have introduced a general definition of context, and of context-centered IR, in the following sections we will focus on personalized IR, i.e. to IR approaches which take advantage of the knowledge represented in a user model, also called user's profile.

III. Modeling the User Context in Personalised IR

In recent years, a great deal of research has addressed the problem of personalizing search, to the aim of taking into account the user context in the process of assessing relevance to user's queries. These research efforts are witnessed both by the numerous publications, and by the existence of conference devoted to personalized approaches to IR, or more generally to IR in context (e.g. the Symposium on Information and Interaction in Context, IIiX [43], the International Conference on User Modeling, Adaptation and Personalization [44], the SIGIR Desktop Search Workshop: Understanding, Supporting, and Evaluating Personal Data Search [45]).

Moreover personalized approaches to IR may be experienced by users through personalized versions of search engines, such as iGoogle, Google Personalized Search (www.google.com/ig).

To personalize search results means to explicitly make use of the user preferences to tailor search results. If for example a query such as "good restaurant in Rome" is formulated by a vegetarian user, the expected results should take this preference into account. To this aim the query evaluation should make explicit use of this information as an additional constraint (besides the query) to estimate document relevance. As another example let us consider a group of users represented by researchers working in an information retrieval lab. If the query "information retrieval" is formulated by the lab director, the query evaluation should produce a different list of documents than the same query formulated by a novice student. In this last example, the user preferences are related to his/her cognitive context, and expertise. The previous simple examples outline that the quality of the search outcome strongly depends on the information beyond the one expressed in a user's query. So the effectiveness of the system strongly depends on the available quantity and quality of information about the user and its preferences. The more accurate the user model is, the more effective the personalized answer can be.

An obvious question rises at this point: how to make this information available to an IR system? To do so three kind of processes should be undertaken: acquisition, representation and updating. The acquisition process is aimed at capturing the information characterizing the user context. The formal representation process is aimed at formally representing the acquired information; this is needed to make it possible that this information be accessed and used by the IRS. The updating process is finalized at learning the changes of the user preferences in time. In the following we shortly discuss each of the above processes. It is clear that the effectiveness of the algorithms which exploit the knowledge of the user context strongly depend on the quality and reliability of the user model (user profile). So the generation of a user model is an important although difficult task.

To capture user's interest two main techniques may be employed: explicit and implicit [17,28]. By the explicit approach the user is asked to be proactive and to directly communicate to the system his/her data and preferences. This can be done by compiling questionnaires, by providing short textual descriptions (to specify topical preferences), and/or by providing a few documents that represents well the user preferences. The texts will be processed by the system to automatically extract their main descriptors. However, an explicit request of information to the user implies to burden the user, and to rely on the user's willingness to specify the required information. This is generally unrealistic. To overcome this problem, several techniques have been proposed in the literature to automatically capture the user's interests, by monitoring the user's actions in the user system interaction, and by inferring from them the user's preferences. The proposed techniques range from click-trough data analysis, query log analysis, desktop information analysis, document display time,

etc [1,11,22,23,30,31,34,37]. The advantage in adopting such techniques is that several sources of knowledge may be considered; the main disadvantage is that automatic processes may be error-prone, as they may introduce noise in the process of identifying the useful information. However, the advantages of using such techniques has revealed much greater than their limitations.

The process of organization and representation of the information obtained by the acquisition phase implies the selection of an appropriate formal language to define the user model. In the literature several representations for the user model have been proposed, ranging from bag of words and vector representations, to graph-based representations, and, more recently, to ontology based representations [8,11,14,16,32,35,40]. The more structured and expressive the formal language is, the more accurate the user model can be. As most current approaches to the definition of user profiles are aimed at defining models based on words or concept features, to the aim of also representing the relations between words/concepts, an external knowledge resource, such as the ODP (Open Directory Project [46], or Wordnet [47]) is required.

An important aspect related to user profiles concerns profile updating; this aspect is generally considered by the research contributions that propose the definition of user models.

IV. Exploiting the user context to enhance search quality

As outlined in section III, the availability of a user model, where the relevant information that characterizes the user context is represented, is necessary to define any process aimed at tailoring, based on this context description, the results proposed as an answer to a user's query. The quality of the personalization process is strongly related to the quality of the user's model, e.g. to its reliability and accuracy.

In the literature several approaches have been proposed, which can be roughly categorized in three main classes [28]:

- approaches to modify/define relevance assessment
- approaches to query modification
- approaches to results re-ranking

Among the approaches belonging to the first category we cite the PageRank based methods, which have proposed modifications of the PageRank algorithm that include user modeling into rank computation, to create personal views of the Web [18,21].

The approaches to results re-ranking are aimed at modifying the ranking score by explicitly matching the user profile against the user query, and then at combining the obtained score with the relevance based score produced by the traditional IRS or search engine. Re-ranking techniques proposed in the literature may differ both in the adopted user model and in the re-ranking strategy. Among the several techniques proposed in the literature we cite [27,31,32].

Query modification techniques are aimed at exploiting the user profile as a knowledge support to select information useful to define more accurate queries via a query expansion or modification technique. Among the techniques proposed in the literature we cite [9,10,26,37].

More recently an interesting problem has been considered to leverage search through a better user knowledge and interaction: this is the problem of visualizing search results in an effective way. One of the biggest problems when using search engines is that, although the information relevant to the user's needs expressed in a query could be probably found in the long ordered list of results, it is quite difficult to locate it. It is in fact well know that users seldom go beyond an analysis of the first two/three pages of search results. An interesting research idea is to enhance results visualization through the knowledge of the user's topical preferences. In [2,4] an approach related to the exploratory search task is proposed, which combines personalized search with a spatial and adaptive visualization of search results

Independently of the decision about how to exploit the knowledge of the user's preferences, an interesting aspect which emerges in context-aware IR is that the availability of a model of context (which may represent both user's preferences, the geographic and social contexts etc.) makes it possible to consider several new dimensions in the relevance assessment process. The birth of Web Search Engines as well as the IR techniques evolution, have implied a shift from topical relevance assessment (which was the only dimension to assess relevance in the first IRSs) to a multi-dimensional relevance assessment, where the considered relevance dimensions encompass topical relevance, page popularity (based on link analysis in web search engines), geographic and temporal dimensions, etc. The availability of a user model (and more generally the availability of more structured context models), make the dimensions available to concur in the process of relevance assessment more numerous. As a consequence, the need of combining the relevance assessments related to each dimensions arise. This problem has been faced so far by adopting simple linear combination schemes, applied independently on the user's preferences over the relevance dimensions. An interesting research direction related to personalized search is to make the user an active player in determining such an aggregation scheme: this could be simply done by making the aggregation dependent on the user's preferences over the single relevance dimensions. In this way, for a same query and a same profile different document rankings can be obtained based on the user's preference over the relevance dimensions. In [12] this approach has been proposed to define user-dependent aggregation schemes defined as linear combinations where weights of relevance dimensions are automatically computed based on the user-specified priority order over the dimensions.

V. The privacy and the evaluation issues

Two important issues that have been addressed in the literature related to personalized IR concern user privacy and the evaluation of the effectiveness of personalized approaches to IR. We start by shortly discussing the privacy issue.

As it has been synthetically discussed in the previous sections, the approaches to personalization strongly rely on user related and user personal information, with the obvious consequent need of preserving users' privacy. In [25,41] very interesting and exhaustive analysis of the privacy issue are presented. As well outlined in these contributions, the user is not incline to make the information that concerns his/her private life available to a centralized system, with the main consequence that often users prefer not to use the personalization facilities. As suggested by the authors, a feasible solution to the privacy issue problem is to design client-side applications.

Systems evaluation is a fundamental activity related to the IR task. The usual approach to evaluate the effectiveness of IRSs is based on the Cranfield paradigm, which is the basic approach undertaken by the TREC (Text Retrieval) Conferences [48]. However, as well outlined in [6], the Cranfield paradigm is not able to accommodate the inherent interaction of users with information systems. The Cranfield evaluation paradigm is in fact based on document relevance assessment on single search results, not suited to interactive information seeking and personalized IR, as it assumes that users are well represented by their queries, and the user's context is ignored. In [36] a good overview of the problem of evaluating the effectiveness of approaches to personalized search is presented. The evaluation of systems that support a personalized access to information encompasses two main aspects, related to the components which play a main role in these systems, i.e. the user model and the personalized search processes. To evaluate a user profile means to assess its quality properties, such as accuracy. With respect to the evaluation of systems' effectiveness, the authors outline in [36] three main approaches undertaken to set up a suited evaluation setting for personalized systems; by the first approach, an attempt to extend the laboratory-based approach to account for the existence of contextual factors were proposed within TREC [5,18]. By the second approach a simulation-based evaluation methodology has been proposed, based on searchers simulations [42]. By the third approach, the one which is most extensively adopted, user-centered evaluations are defined, based on user studies, with the involvement of real users who undertake qualitative system's evaluations [24]. Evaluation is a quite important issue that deserves special attention, and which still needs important efforts to be applied to context-based IR applications.

VI. Conclusions

To conclude this short overview of some main issues related to personalized IR, we want to mention a promising research direction which aims at exploiting the users' social context to produce more effective results in Web Search [33,39]. This is made possible by recent applications and technologies related to the so called Social Web, aimed at making the user active in both content generation and sharing. In [33] an approach to collaborative Web Search has been recently proposed, which based on the search behavior of a community of like-minded users is aimed to adapt results of conventional search engines to the community preferences.

References

[1] E. Agichtein, E. Brill, S. Dumais, R. Ragno, "Learning user interaction models for predicting Web search preferences", in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 3–10.

[2] J. Ahn and P. Brusilovsky, "From User Query to User Model and Back: Adaptive Relevance-Based Visualization for Information Foraging", in *Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence*, Silicon Valley, CA, USA, 2007, pp. 706–712.

[3] J. Ahn, P. Brusilovsky, D. He, J. Grady, Q. Li, "Personalized Web Exploration with Task Modles", *in Proc. of the17th international conference on World Wide Web (WWW 2008)*, Bejing, China, 2008, pp. 1-10.

[4] J. Ahn and P. Brusilovsky. "Visual interaction for personalized information retrieval", in Third Workshop on Human-Computer Interaction and Information Retrieval (HCIR 2009), Washington, DC, USA, Oct 2009.

[5] J. Allan J, "Hard track overview in TREC 2003 high accuracy retrieval from documents", in Proc. of the 12th text retrieval conference (TREC-12), National Institute of Standards and Technology, NIST special publication, pp 24–37.

[6] N.J. Belkin, " Some(what) Grand Challenges for Information Retrieval, in *SIGIR Forum*, Vol. 42, No. 1. (2008), pp. 47-54.

[7] P. Brusilovsky, A. Kobsa, W. Nejdl, eds. *The adaptive web: methods and strategies of web personalization*, Springer, LNCS 4321, 2007.

[8] S. Calegari, G. Pasi, "Ontology-Based Information Behaviour to Improve Web Search", in *Future Internet*, 4, 2010, pp. 533-558.

[9] P. Chirita, C. Firan, W. Nejdl, "Summarizing local context to personalize global Web search", in *Proc. of the Annual International Conference on Information and Knowledge Management, CIKM 2006*, pp. 287–296.

[10] P. Chirita, C.Firan, W. Nejdl, "Personalised Query Expansion for the Web", in *Proc. of the ACM SIGIR Conference 2007*, pp. 287–296.

[11] M. Claypool, D. Brown, P. Le, M. Waseda, " Inferring user interest", in *IEEE Intern.Comput.* 32–39, 2001.

[12] C. da Costa Pereira, M. Dragoni, G. Pasi, "Multidimensional Relevance: A New Aggregation Criterion", in Proc. Of the European Conference on Information Retrieval (ECIR 2009), pp. 264-275.

[13] M. Daoud, L. Tamine-Lechani, M. Boughanem, "Towards a graph-based user profile modeling for a session-based personalized search", in *Knowl. Inf. Syst.*, 21 (3), 2009, pp. 365-398.

[14] M. Daoud, L. Tamine, M. Boughanem, "A Personalized Graph-Based Document Ranking Model Using a Semantic User Profile", *in Proc. of UMAP 2010*, pp. 171-182.

[15] Z. Dou, R. Song, J.R. Wen, "A large-scale evaluation and analysis of personalized search strategies", in *Proc. of the International World Wide Web Conference*, 2007, pp. 581–590.

[16] S. Gauch, J. Chaffee, A. Pretschner, "Ontology-Based perosnalized search and browsing", in *Web Intelligence and Agent Systems: an International Journal*, 2003, pp. 219-234.

[17] S. Gauch, M. Speretta, A. Chandramouli, A. Micarelli, "User Profiles for Personalized Information Access", The Adaptive Web 2007, pp. 54-89

[18] T. Havelivala, Topic-sensitive PageRank, ", in Proc. 12th International orld Wide *Web Conference* (WWW 2003), Hawai, 2002.

[19] D. Harman, overview of the 4th text retrieval conference (TREC-4). In *Proc. Of the 4th text retrieval conference (TREC-4)*, National Institute of Standards and Technology, NIST special publication, pp. 1-24.

[20] P. Ingwersen, and K. Järvelin, *The Turn: Integration of Information Seeking and Retrieval in Context*, Springer-Verlag Eds, 2005

[21] G. Jeh and J. Widom, "Scaling Personalized Web Search", *in Proc. 12th International World Wide Web Conference* (WWW 2003), Budapest, Hungary, 2003, pp. 271279.

[22] T. Joachims, L. Granka, B. Pang, H. Hembrooke, G. Gay, "Accurately interpreting clickthrough data as implicit feedback", in *Proc. of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2005, pp. 154–161.

[23] D. Kelly and J. Teevan, "Implicit feedback for inferring user preference: A bibliography", in *SIGIR Forum 37*, 2, 18–28, 2003.

[24] D. Kelly, "Methods for evaluating interactive information retrieval systems with users", in Foundations and Trends in Information Retrieval, DOI: 10.1561/15000000123, (1-2), 2009, pp. 1-224..
[25] A. Kobsa, "Privacy Enhanced Personalization", in P. Brusilovsky, A. Kobsa, W. Nejdl, eds. *The Adaptive Web: Methods and Strategies of Web Personalization. Berlin, Heidelberg*, New York, Springer-Verlag, 2007, 628-670.
[26] C Liu, C. Yu, and W. Meng, "Personalized Web Search For Improving Retrieval Effectiveness", in *IEEE Transactions on Knowledge and Data Engineering*, 16(1), January 2004, pp. 28-40.
[27] Z. Ma, G. Pant, and O. Sheng," Interest-based personalized search", in *ACM Transaction on Information Systems, 25*, 5, 2007.
[28] A. Micarelli, F. Gasparetti, F. Sciarrone, S. Gauch, "Personalized Search on theWorld Wide Web", in Lecture Notes in Computer Science, 2007, 4321, pp.195–230.
[29] J. Pitkow, H. Schutze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel, "Personalized search", in *Communications of the ACM, 45*, 9, 2002, pp. 50–55.
[30] X. Shen, B. Tan, and C.X. Zhai, "Context sensitive information retrieval using implicit feedback", in *Proc. of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2005, pp. 43-50.
[31] X. Shen, B. Tan, and C.X. Zhai, "Implicit user modeling for personalized search", in *Proc. of the International Conference on Information and Knowledge Management*, CIKM 2005, pp. 824–831.
[32] A. Sieg, B. Mobasher, R. Burke ,"Web search personalization with ontological user profiles", in *Proc. of the International Conference on Information and Knowledge Management* , CIKM 2007, pp. 525-534.
[33] B. Smyth, "A Community-Based Approach to Personalizing Web Search", in IEEE Computer, 40(8), 2007, pp.42-50.
[34] M. Speretta, S. Gauch: Personalized Search Based on User Search Histories", in *Proc. of the IEEE/WIC/ACM International conference on Web Intelligence, 2005*, pp. 622-628.
[35] M. Speretta, S. Gauch, "Miology: a Web Application for Organizing Personal Domain Ontologies", in *Proc. of the International Conference on Information, Process and Knowledge Management*, 2009, pp. 159-161.
[36] L. Tamine-Lechani, M. Boughanem, M. Daoud, "Evaluation of contextual Information Retrieval effectiveness: overview of issues and research", *Knowledge Information Systems*, 24, 2010, pp. 1-34.
[37] J. Teevan, S. Dumais, and E. Horvitz, " Personalizing search via automated analysis of interests and activities", in *Proc. of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2005, pp. 449–456.
[38] J. Teevan, S.T. Dumais, and D. J. Liebling, "Personalize or not to personalize: Modeling queries with variation in user intent", in *Proc. of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2008, pp. 163–170.
[39] J. Teevan, M. Morris, S. Bush, "Discovering and using groups to improve personalized search". *In Proceedings of the ACM International Conference on Web Search and Data Mining*, 2009, pp. 15–24.
[40] J. Trajkova, S. Gauch, "Improving Ontology-Based User Profiles", in *Proc. of the. RIAO 2004 Conference*, pp. 380-390
[41] E. Volokh, "Personalization and Privacy", in *Communications of the ACM*, 43(8), 2000.
[42] R. W. White, I.Ruthven, J. Jose, C. van Rijsbergen, "Evaluating implicit feedback models using seracher simulations", in *ACM Transactions on Information Systems*, 22(3), 2005, pp. 325-361.
[43] http://www.iva.dk/IIiX/
[44] http://www.hawaii.edu/UMAP2010/
[45] http://www.cdvp.dcu.ie/DS2010/
[46] http://dmoz.org/
[47] http://wordnet.princeton.edu/
[48] http://trec.nist.gov/