"ASSISTED BROWSING" APPROACH FOR USER PROFILE CAPTURE IN PERSONALISATION SYSTEMS

ÁLVARO RAMOS, PEDRO CONCEJERO, PAULO VILLEGAS
Telefónica Investigación y Desarrollo
Emilio Vargas, 6. E-28043 MADRID SPAIN
arc@tid.es

AND

JÉRÔME PICAULT
Motorola Labs
Parc Les Algorithmes, 91193 Gif-sur-Yvette, France
jerome.picault@motorola.com

Abstract. This paper aims to make a proposal for overcoming some of the user preference capture process limitations, by means of a mechanism of "Assisted Browsing" for the initial moments of the user interaction. The approach we present introduces a mechanism for users to browse, in a guided way (what we call "assisted browsing"), among subsets of the real content so that they can express their preferences more effectively, and at the same time, the system can capture data related to actual user decisions ("Yes, more like this", "No, take me elsewhere") with respect to content, instead or mere opinions about the meaning of the categories. The concept of "Assisted Browsing" has been taken from some systems appeared in late 90s and which can be found in some services for providing catalogues in e-commerce. However, our approach uses some procedures coming from Psychometrics computerized tests (or CAT, Computer Assisted Testing), which are new in the personalisation field. This paper reports on the ongoing research, including some details on the initial specification of the algorithms.

1. Introduction

Personalised systems are starting to be more and more present in consumers' lives. Systems such as Amazon, TiVo, StumbleUpon, etc. are already extremely useful: it has been reported (Wall Street Journal, 2007) that such recommendation systems generally increase sales by more than 10%. Recent research on user-generated content consumption (Burke, 2002) has concluded that improving the search and filtering mechanisms could increase the total views by as much as 45%. Although online providers are relying on increasingly sophisticated personalisation techniques to help customers to choose the best products or simply to increase the volume or speed of sales, there are still open challenges; in particular, the interaction with the user is often neglected, which may result in a lack of understanding of recommendations or impossibility for the user to correct the system.

According to ChoiceStream Technology brief, there are very clear user requirements for personalisation systems: relevance and precision of recommendations; real-time (immediate actions on consumers); and minimal (but not lack of) involvement from consumers. Thus, Bonnefoy et. al. (2007) report that users request to provide explicit feedback to the system. A profile describing the user's preferences is an essential component of any personalisation system. Depending on the kind of recommender which is used in the personalised system, it may be very simple (list of choices or ratings) or more complex (summary of users' preferences described by concepts that belong to a taxonomy or ontology). The accuracy of the user model has a direct impact on the effectiveness of the personalisation system.

The main risk of an inappropriate user profile is that the user may be presented with content that is not interesting and later find out other content in the system that matches his or her interests. If these errors occur too often, then user will quickly lose trust in the system's ability to choose the appropriate content. In a typical personalised system, user preferences can be acquired explicitly from user responses, usually to a form or questionnaire, or implicitly through a learning process. Implicit techniques for acquiring user preferences are generally preferred by developers or researchers but this feeling is not shared by users who "seem neither to trust nor like a system that would silently learn preferences on their behalf" (Bonnefoy et. al., 2007). For instance a new user of a movie recommendations system can rate very highly Hitchcock's "The birds" without knowing that this movie is within the category "terror", and with no other terror film rated, his rating for that category in the automatically generated profile can result very high. Only by a close and regular review of the profile can the user notice and correct these errors. But users almost always consider entering preferences as a tedious task (Jameson, 2004), which they often skip; this may result in bad recommendations and distrust from the users about the personalised system. In this paper, we focus on explicit acquisition of preferences, but not with the traditional form with a long list of categories. We start by evoking some the issues of explicit acquisition of user preferences and then we present our approach to user preference capture to make this process more attractive.

2. Issues in acquiring user preferences

The explicit acquisition of user profile needs an adapted and fit-for-purpose interaction mechanism to help the user deal with this "tedious" phase of personalisation initialisation. In this section, we summarize briefly the existing approaches and their limitations.

2.1.ACQUISITION OF PREFERENCES VIA A HIERARCHY OF CONCEPTS

In the case of explicit acquisition, user preferences are often described thanks to a tree of interest categories where the user can click and choose categories he is interested in, such as (Baudisch, 2005; Duke and van der Meer, 2002). The category list or tree is defined at design time, for example according to a user study, or by relying on knowledge experts. As Hurwitz (2006) puts it: "One assumption of this approach is that user content preferences can be predicted from their preferences for content categories. (...) However, this assumption has not been empirically tested across a variety of content domains".

In this case, several problems may arise:

- **Interface issues**: if the designed list of interests is comprehensive, explicit acquisition through a user interface is tedious because of the complexity of the interests' list; on the other end, sometimes a compromise is made with the efficiency of the user profile acquisition and the interests' list is limited to the most popular categories. Thus, attempting to increase accuracy by adding more category terms for use in the personal profile and in the metadata will increase the work required by users to set up their profiles. It could also, paradoxically, increase the likelihood of error, since there are more opportunities for erroneous mismatches between the metadata and the terms stored in the personal profile."
- **Understanding categories**: this approach relies on the presumption that the system for categorizing content reflects the user's mental representation of that content. Usually the taxonomies or ontologies are designed by expert people and only reflect their 'viewpoints' (Ribière, 1999), not necessarily the perception of the end-user. If preferences are defined based on concepts defined in taxonomies, they may not provide the right granularity for any given end-user (too specific, too generic or too heterogeneous), thus causing confusion to a potential user to express preference for it. An example taken from the IPTC NewsCodes is "Human Interest", which includes: Animals, Curiosities, People, Advice, Celebrities, Mysteries, Society, Awards and Prizes, Imperial, Royal Matters. If you have a close look at the extremely long list of labels for subcategories you will always find something of your interest, something you do not know what it means, and of course uninteresting items. Therefore this approach can produce expansive hierarchies that are long, difficult to understand and in some cases, not relevant for the task of expressing preferences.

- **Inconsistencies between interest for the category and interest for the content for a particular domain:** Hurwitz (2006) found that "interest in content categories may not necessarily be a reliable indicator of interest in the content", particularly for typical content like Business, Health and Entertainment domains.

2.2. ACQUISITION OF PREFERENCES VIA A LIST OF CONTENT ITEMS

Some other systems ask the user to fill indirectly a user profile, not by letting them describe directly themselves through high-level concepts, but by letting the user choose (or asking him to choose) content items s/he likes, or to rate browsed or played content; then the system can infer from these inputs additional user preferences, for example by associating the user to a pre-defined stereotype.

This approach makes the process simpler for the user. For example, in PTVplus (Smyth and Cotter, 2001), preliminary profile information is collected from the user at registration time to bootstrap the personalisation process (the user can rate content from the program guide positively or negatively). But in such a system, the user does not know how much content s/he has to evaluate (i.e. when the acquisition phase terminates), and it appears that in practice, when a list of content is presented to users, they tend to rate mainly content they like, while ignoring content they dislike instead of rating it negatively. Choicestream, a technology provider of many personalised systems, also uses a similar approach: users can fill out a short 'jumpstart' questionnaire, which asks just a few, well-designed questions. It uses preferences of groups of similar users to inform individual profiles. The user is always associated with one or more cohorts of users. As the system learns more about the preferences of a user's cohort(s), it updates the user's preference profile based on that information accordingly.

The approach we present is an extension / improvement of this second family of preference acquisition, by introducing a mechanism for users to browse, in a guided way (what we call "assisted browsing"), among subsets of the real content so that they can express their preferences more effectively, and at the same time, the system can capture data related to actual user decisions with respect to content, instead or mere opinions about the meaning of the categories. The challenge is to quickly collect as much preference data as possible so that users can begin working toward their goals. The concept of "Assisted Browsing" has been taken from some systems appeared in late 90s (Andersn et. al., 1997) and which can be found in some services for providing catalogues in e-commerce (Burke, 2002). Another area which uses similar iterative approaches is preference elicitation for decision-making help systems (Chajewska et. al., 2000), though the models they are based on are normative rather than descriptive, or differential, as appropriate for personalisation. However, our approach uses some procedures coming from Psychometrics computerized tests, which are new in the personalisation field.

3. Psychometric models for preference measurement

In Psychology the measurement of preferences and attitudes has a long tradition. For instance, Thurstone in 1927 developed the method of pairwise comparisons as an approach to measuring perceived intensity of physical stimuli, attitudes, preferences, choices, and values, using simple models known as "Law of Comparative Judgment". More advanced models were proposed later, based on these. One is the Bradley-Terry-Luce (BTL) model (Luce, Individual Choice Behavior, 1959), which was often applied to pairwise comparison data to scale preferences.

There is another evolution of this psychological measurement approach, known as Item response theory (IRT) that has become the standard for psychological measurement in education. IRT is the generic name for a variety of formal models which have in common that they establish a mathematical relationship between the response to a particular "item" (e.g., a question in a verbal comprehension test) and a level of an internal trait of the person. That is why they were also called "latent trait models". They allow for the simultaneous scaling of both subjects (test respondents) and the latent trait. The latent trait can be any aspect or construct of human behavior in which individual differences are to be measured, as academic performance, personality constructs, attitudes, and, of course, interests. This is usually referred to as θ. The most widely applied model is the logistic two-parameter model, which for an item (i.e., question of a test) *i* is (Lord and Novick, 1968):

$$p_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta_a - b_i)}} \tag{1}$$

where a_i is the discrimination parameter of the item, b_i is the parameter difficulty of the item, and D is a constant as a scaling factor to approximate the probabilities to those of a normal distribution.

This model allows for a full description of an item with respect to the latent trait in which is called "Item Information Curve", or IIC, which has a familiar S-shape, whose slope, or change rate, is defined by b_i. Theoretically the difficulty parameter describes where the item is located in the scale of the latent trait, or also the position of the item in the measurement scale of the latent trait. However all the theoretical advantages of IRT models, they impose strict assumptions that have to be proved empirically for a particular application domain. The most important of them is unidimensionality of the latent-trait. Though never perfect in almost any field, it is important that the response to an item is explained by a single dominant factor. Another assumption is local independence.

Estimation of the model is relatively complex, using maximum likelihood procedures, but plenty of software exists for this purpose. However, it is well known that good estimates require many points of the curve, thus requiring many subjects to make some tasks to extract the data. IRT solved, theoretically and in practice, subject-adaptive tests, and, crucially for us, computer-assisted testing. Opposite to traditional tests, which have a fixed length, IRT allows for a dynamic and iterative approximation to the subject score, and can reduce considerably of a test.

An important technical issue for this is the threshold, or termination criterion. Several approaches exist, whose description is beyond the scope of this paper, but which can be grouped as IRT-based confidence intervals, the sequential probability ratio test (SPRT), and decision theory. The former is the one requiring larger samples of subjects, and the second is the most commonly used method (Thompson, 2007), requiring the estimation of two latent-trait levels in a controlled situation, which defines an “indifference region”. The wider this region the shorter the test, but at the same time, the potentially less accurate.

4. What mechanism does the “assisted browsing” wizard follow?

The wizard mechanisms mimic the procedures implemented in Psychometrics Computer-Adaptive Tests or CAT (Wikipedia.org). A CAT is composed of several components, of which the one used as input is the “Calibrated item pool”. This is critical for the appropriate performance of the rest. This is obtained from psychometric studies that estimate the parameters as defined above for pools of items (tasks that the examinee has to solve), and performance levels according to examinee profile (such as educational level). These items are in our context the content elements that will be presented to the user. The difficulty can be extrapolated to a “rarity” parameter (or probability that a particular user would show interest for that item). This is the basis for the content set that forms one essential part of the wizard. Many possibilities exist, from a single wizard per category to an adapted one based on previous choices or on the identification of the user as belonging to one user typology.

The wizard starts with a pre-defined starting point. After the user has responded to an item (one of the contents presented), the wizard computes a score based on the logistic 2-parameter model. The wizard also computes the “Termination criterion”, and stops the procedure, for that particular category of items, or content belonging to a particular type of user, when the cutoff point is found. The estimate of the latent trait is then stored as the "interest" of the user for that category or stereotype. Figure 1 below depicts the proposed mechanism.

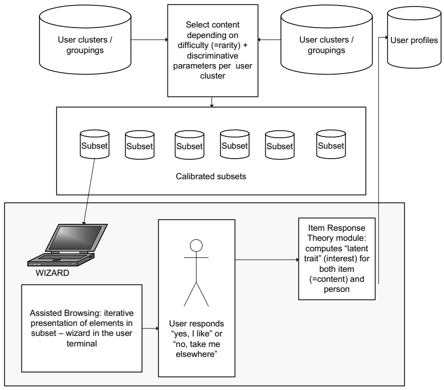

Figure 1. Components of the proposed "Assisted Browsing" mechanism.

Of course the termination for a particular content set can also be forced by the user, simply stating "I am not interested in any of these" and going to other content sets. A pre-defined maximum number of iterations for a single content set can also be defined.

5. Conclusion and future work

MESH is considering the application of this approach in the news context, and testing between the different scoring models. Though literature about its use in applied contexts (e.g., CAT software) is extensive, this is not the case in personalisation systems.

Theoretically, in our opinion, there is no problem in using IRT for preference measurement. In practice, however, there are important challenges that have to be tested empirically, most important of them to know if the performance of a personalisation system which uses these psychometric preference measures is significantly better than other more traditional, and clearly simple, approaches.

Acknowledgements

The research leading to this document has received funding from the European Community's Sixth Framework Programme (IST-FP6-027685 MESH).

References

Anderson, R.E. Burke, R. D., Hammond, K. J., Young, B. C.: 1997, The FindMe Approach to Assisted Browsing, *IEEE Expert*, jul. aug. 1997, pp. 32-40.

Baudisch, P.: 2005, The Profile Editor: Designing a direct manipulative tool for assembling profiles, *Proc. of Fifth DELOS Workshop on Filtering and Collaborative Filtering*, pp. 11-17.

Bonnefoy, D.; Bouzid, M.; Lhuillier, N.; Mercer, K.: 2007, "More like this" or "Not for me" – Delivering Personalised Recommendations in Multi-User Environments, *in Proc.of 11th International Conference, UM2007*, LNAI 4511, Springer Verlag, pp. 87-96.

Burke, R.: 2002, Interactive Critiquing for Catalog Navigation in E-Commerce, *Artificial Intelligence Review*, 18, 245-267.

Cha, M.; Kwaky, H.; Rodríguez, P. *et al.*: 2007, I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System, *IMC'07,* October 24-26, 2007, San Diego, California, USA.

Chajewska, U., Koller, D. and Parr, R.: 2000, Making rational decisions using adaptive utility elicitation, in *Seventeenth National Conference on Artificial Intelligence (AAAI'00),* pp. 363-369.

ChoiceStream technology brief:
http://www.choicestream.com/techbrief/ChoiceStream_TechBrief.pdf

Duke, A. and van der Meer, J.: 2002, User Profile Construction, On-To-Knowledge project: Content-driven Knowledge management Tools through Evolving Ontologies, http://www.aduna-software.org/home/papers/OTK-del14.pdf

Hurwitz, J. B.: 2006, Empirical Evaluation of Content-Based Filtering for Personalization, *20th International Symposium on Human Factors in Telecommunication*. Sophia-Antipolis, France, 20-23 March, 2006.

Jameson, A.: 2004, More than the sum of its members: Challenges for group recommender systems, *Proceedings of the International Working Conference on Advanced Visual Interfaces,* pp. 48-54.

Lord, F. M.; Novick, M. R.: 1968, Statistical theories of mental test scores, Reading, MA: Addison-Wesley.

McCarthy, K., Reilly, J., McGinty, L., Smyth, B.: 2005, *An Analysis of Critique Diversity in Case-Based Recommendation*, American Association for Artificial Intelligence.

McCarthy, K., Reilly, J., Smyth, B. McGinty, L.: 2005, Generating Diverse Compound Critiques, *Artificial Intelligence Review,* 24, 339-357.

Ribière, M. : 1999, *Représentation et gestion de multiples points de vue dans le formalisme des graphes conceptuels,* PhD thesis, University of Nice, France.

Smyth, B. and Cotter P.: 2001, Personalized Electronic Program Guides for Digital TV, *AI Magazine,* **22** (2), pp 89-98.

Wall Street Journal article (07/31/07): We Know What You Ought To Be Watching This Summer.

International Press Telecommunications Council, IPTC NewsCodes, http://www.iptc.org/NewsCodes/

Wikipedia.org, [online] http://en.wikipedia.org/wiki/Computer-adaptive_test

Thompson, N. A.: 2007, A Practitioner's guide for variable-length computerized classification testing, *Practical Assessment, Research & Evaluation*., 12 (1), January 2007.