Interactive Preference Elicitation for Scientific and Cultural Recommendations

Eduardo Veas[1,2] **and Cecilia di Sciascio**[2]
[1] Information and Communications Technologies, National University of Cuyo
[2] Knowledge Visualization, Know-Center GmbH
eduveas@gmail.com, cdissciascio@know-center.at

Abstract

This paper presents a visual interface developed on the basis of control and transparency to elicit preferences in the scientific and cultural domain. Preference elicitation is a recognized challenge in user modeling for personalized recommender systems. The amount of feedback the user is willing to provide depends on how trustworthy the system seems to be and how invasive the elicitation process is. Our approach ranks a collection of items with a controllable text analytics model. It integrates control with the ranking and uses it as implicit preference for content based recommendations.

1 Introduction

A recommender system (RS) depends on a model of a user to be accurate. To build a model of the user, behavioral recommenders collect preferences from browsing and purchasing history, whereas rating recommenders require a user to rate a set of items to state their preferences (implicit and explicit methods respectively) [Pu *et al.*, 2011]. Preference elicitation is fundamental for the whole operational lifecycle of a RS: it affects the recommendations for a new user and also those of the whole system community, given what the RS learns from each new user [Cremonesi *et al.*, 2012]. Whichever method is chosen, preference elicitation represents an added effort, which may be willingly avoided to the detriment of user satisfaction. The amount of feedback the user is willing to provide is a tradeoff between system aspects and personal characteristics, for example privacy vs recommendation quality [Knijnenburg *et al.*, 2012].

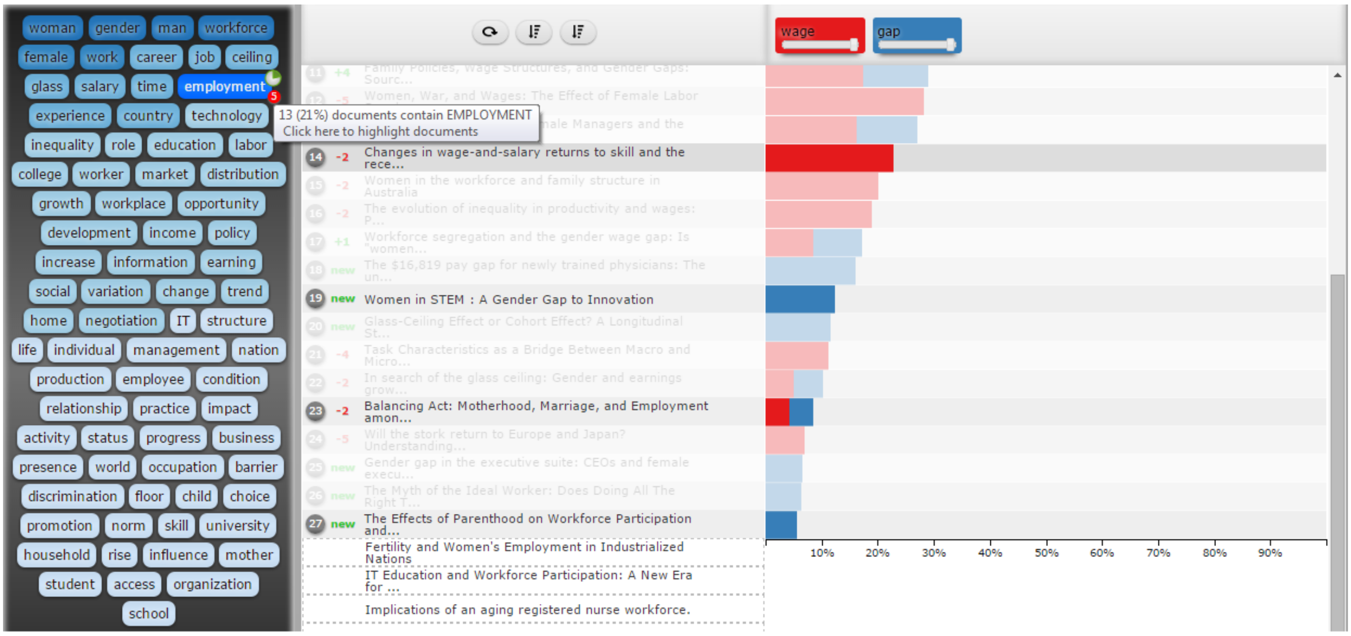
In their seminal work, Swearingen et al. pointed out one challenge: the recommender has to convince the user to try the recommended items [Swearingen and Sinha, 2001]. To do so, the recommendation algorithm has to propose items effectively, but also the interfaces must deliver recommendations in a way that can be compared and explained [Ricci *et al.*, 2011]. The willingness to provide feedback is directly related to the overall perception and satisfaction the user has of the RS [Knijnenburg *et al.*, 2012]. Explanation interfaces increase confidence in the system (trust) by explaining how the system works (transparency) [Tintarev and Masthoff, 2012] and allowing users to tell the system when it is wrong (scrutability) [Kay, 2006]. Hence, to warrant increased user involvement the RS has to justify recommendations and let the user customize their generation. Transparency and controllability are key facilities of a self-explanatory RS that promote trust and satisfaction [Tintarev and Masthoff, 2012]

Our work is set in the scientific and cultural domain. In this frame, users are most often engaged in exploration and production tasks that involve gathering and organizing large collections in preparatory steps (e.g., for writing, preparing a lecture or presentation). A federated system (FS) compiles scientific documents or electronic cultural content (images) upon an explicit or implicit query, with little control over the way results are generated. Content takes the form of text document surrogates comprising title and abstract. They also include minimal additional metadata, like creator, URL, provider and year of publication.

This paper introduces a visual tool to support exploration of scientific and cultural collections. The approach includes a metaphor to represent a set of documents, with which the user interacts to understand and define themes of interest. The contribution of this work is the interactive personalization feature that, instead of presenting a static ranked list, allows users to dynamically re-sort the document set in the visual representation and re-calculate relevance scores with regards to the own interests. The visual interface employs controllable methods to represents their results in a transparent manner which, rather than adding effort, reduces complexity of the overall task.

2 The Approach

The proposed approach was designed to quickly reorganize a large collection in terms of its relevance to a set of keywords expressing the choice of topic. In a nutshell, the goal is to interactively discover the topics in a collection, building the knowledge in the user. But, instead of trying to infer a hidden topic structure fully automatically (as in [Blei, 2012]), we propose an interactive approach, which works as a conversation between the user and the RS to build a *personalized theme structure*. Controllability and transparency are crucial for the user to understand how a topic came about from their personal exploration. The challenge for the interface is to clearly explain the recommendation process, and for the analytics method to reduce the computational problem to interactive terms.



13 (21%) documents contain EMPLOYMENT
Click here to highlight documents

11 +4 Family Policies, Wage Structures, and Gender Gaps: Sourc...
12 -5 Women, War, and Wages: The Effect of Female Labor
…male Managers and the
14 -2 Changes in wage-and-salary returns to skill and the rece...
15 -2 Women in the workforce and family structure in Australia
16 -2 The evolution of inequality in productivity and wages: P...
17 +1 Workforce segregation and the gender wage gap: Is "women...
18 new The $16,819 pay gap for newly trained physicians: The un...
19 new Women in STEM : A Gender Gap to Innovation
20 new Glass-Ceiling Effect or Cohort Effect? A Longitudinal St...
21 -4 Task Characteristics as a Bridge Between Macro and Micro...
22 -2 In search of the glass ceiling: Gender and earnings grow...
23 -2 Balancing Act: Motherhood, Marriage, and Employment amon...
24 -5 Will the stork return to Europe and Japan? Understanding...
25 new Gender gap in the executive suite: CEOs and female execu...
26 new The Myth of the Ideal Worker: Does Doing All The Right T...
27 new The Effects of Parenthood on Workforce Participation and...
Fertility and Women's Employment in Industrialized Nations
IT Education and Workforce Participation: A New Era for ...
Implications of an aging registered nurse workforce.

Figure 1: (Left) TagBox. A summary of the collection contents as a bag of words. (Right) The RankView is updated as two terms have been selected. As the user points at a third keyword (employment), a hint shows which documents would be affected by picking it (3 highlighted documents).

2.1 Visual Interface

To search and explore documents based on the themes that run through them, we build an interface that allows the user to establish a conversation with the RS. Two main parts of the interface comprise the topic summary and the recommendation pane. The topic summary is built from keywords extracted from the whole collection. Keywords are presented in a Tag Box, organized and encoded in terms of their frequency of occurrence in the collection (tf-idf), see Fig. 1. The recommendation list initially shows the unranked collection.

As the user interacts with the contents choosing words to express her information needs, the recommendation list is ranked on-the-fly (see Fig. 1). The RankView shows the contribution each keyword has on the overall score of a document. With a slider, the user can assign a weight to a keyword and modify its contribution to the score. Furthermore, the TagBox and RankView illustrate the possible effect of user actions in a quick overview: mouse over a keyword in the TagBox shows a micro-chart with the proportion of documents affected and the RankView highlights those documents in view that would be affected by choosing the keyword.

It is important to note that the user is aware and in control of the ranking and organization of the document collection at all times. With the visual interface, the user describes her information needs and chooses documents from the collection that better reflect those needs. Chosen items can be assigned to a collection. The act of choosing an item is considered an expression of preference. With the collection, the system stores keywords and score of each document. Although this feedback is not yet incorporated in our ranking approach, we analyze its effects with a user study and outline future directions to integrate this additional information in the system.

2.2 Text Analytics and Ranking

Keyword extraction plays two roles: it summarizes the topics in the collection, and it also provides the basis for the fast ranking of documents. Preprocessing involves part-of-speech tagging, singularizing plural nouns, and stemming with a Porter Stemmer. Resulting terms form a document vector, which also constitutes its index. Subsequently, individual terms are scored with TF-IDF (term frequency - inverse document frequency). It how important a term is to a document in a collection, as the coefficient between its frequency in a document and the logarithm of the times it is repeated in the collection of documents. The more frequent a term is in a document and the fewer times it appears in the corpora, the higher its score will be. TF-IDF scored terms are added to the metadata of each document. To provide an overview of the contents in the collection, keywords from all documents are collected in a global set of keywords. Global keywords are sorted by the accumulated document frequency (DF), calculated as the number of documents in which a keyword appears - regardless of the frequency within the documents.

Quick exploration of content depends on quickly re-sorting the documents according the information needs of the user, expressed with a query built from a subset of the global keyword collection. We assume that some keywords are more important to the topic model than others and allow the user to assign weights to them.

The documents in the set are then ranked and sorted as follows. Given a set of documents $D = d_1, ..., d_n$, a set of keywords: $K = k_1, ..., k_m$ and a set of selected keywords: $T = t_1, ..., t_p, T \subseteq K$; the overall score for document d_i is calculated as the sum of the weighted scores of its keywords matching selected keywords:

$$s_{d_i} = \sum_{j=1}^{p} w_{t_j} \times m_{l_i t_j},$$

Where w_{t_j} is the weight assigned by the user to the selected keyword t_j, such that $\forall j : 0 \leq w_{t_j} \leq 1$; and $m_{d_i t_j}$ is the tf-idf score for keyword t_j in document d_i. D is next sorted by overall score using the quicksort algorithm. Documents in D are now elements of sequence Q with order determined by:

$$Q = (q_i)_{i=1}^{n}, q_i, q_{i+1} \in D \wedge s_{q_i} \geq s_{q_{i+1}}.$$

Finally, the ranking position is calculated in such a way that items with equivalent overall score share the same position. The position for a sorted document q_i is calculated as

$$r_{q_i} = \begin{cases} 1 & \text{if } i = 0 \\ r_{q_{i-1}} & \text{if } s_{q_i} = s_{q_{i-1}} \\ r_{q_{i-1}} + |C| & \text{if } s_{q_i} \text{ ¡ } s_{q_{i-1}} \end{cases}$$

Where $C = q_j / s_{q_j} = s_{q_{j-1}}, 0 \leq j \leq i$ represents the set of all the items with immediate superior overall score than q_i.

The current approach employs a term-frequency-based scheme to compute document scores, as it is more appropriate to compute and highlight individual term contributions than a single similarity measure.

3 Experimental Setup

We performed a preliminary study to determine if controllability and transparency increase the complexity and pose an extra effort in the task of building topic oriented document collections. Thus, participants had the task to "gather relevant items" using our tool (U), or using a recommendation list (L) with usual tools (keyword search).We chose two variations of size of the dataset in terms of item count S(30), L(60).

3.1 Method

The study was structured as a repeated measures design, with four iterations of the same tasks, each with a different combination of the independent variables (e.g., US-LL-UL-LS). To counter the effects of habituation, we used four topics covering a spectrum of cultural, technical and scientific content: women in the workforce (WW), robotics (RO), augmented reality (AR), circular economy (CE). Each of these topics has a well defined wikipedia page, which was used as seed to retrieve a collection from a federated system. The system creates a query from the text of the page and forwards it to a number of content providers. The result is a joint list of items from each provider. The federated system cannot establish how relevant the items are. Furthermore, the resulting collection refers to the whole text, but there is no indication of subtopics. We collected sets of 60 and 30 items as static datasets for each topic. We simulated the proposed scenario of reorganizing the collection by choosing subtopics for each task in the study. The combinations were randomized and assigned using a balanced Latin Square.

Each condition had two fundamental tasks: find items most relevant to a set of given keywords, find items most relevant to a short text. In the former, participants were given the keywords and they just had to explore the collection. There were two iterations of this task for each condition. The short text task required participants to come up with the keywords describing the topic by themselves.

Twenty four (24) participants took part in the study (*11* fem., *13* m., between 22 and 37 years old). They were recruited from the medical university and from computer science university graduate population. None of them is majoring in the topic areas selected for the study.

Procedure

A study session started with an intro video, which explained the functionality of the tool. Each participant got exactly the same instructions. There was a short training session on a dummy dataset to let participants familiarize with the tool. Thereafter, the first condition started. The system showed a short text to introduce the topic. After reading the text, participants pressed start, opening the interface for the first task. At the beginning of the task, the items in the collection were ordered randomly, ensuring that an item would not appear in the same position again. The instructions for the task were shown in the upper part of the screen. In all conditions participants were able to collect items and inspect their collections. In the (L) condition the main interface was a list of items, whereas the (U) condition used the proposed interface. Participants had to click the *finished* button to conclude the task. It was possible to finish without collecting all items. After each condition, participants had to fill a NASA TLX questionnaire to assess cognitive load, performance and effort among others.

The procedure was repeated for each of the four iterations. Thereafter participants were interviewed for comments.

3.2 Results

NASA TLX data were analyzed using a repeated measures ANOVA with independent variables tool, and dataset size. Post-hoc effects were computed using Bonferroni corrected pairwise comparisons. The two by two experimental design ensures that sphericity is necessarily met. A repeated measures ANOVA revealed a significant effect of tool on perceived workload F(1,23)=35.254, $p < 0.01, \epsilon = 0.18$. A Post-hoc paired-samples t-test revealed a significantly lower workload when using uRank ($p < 0.001$). Further, repeated measures ANOVA in each dimension of the workload measure showed significant effects of tool in all dimensions as shown in Table 1.

To test the proposed recommender, we gathered and compared for each topic (WW, RO, AR, CE) the most popular items collected using the list (L-MP) and our approach (U-MP) with the scores received by our ranking algorithm (U).

Table 1: Complexity: people found our tool incurs significantly lower workload in all dimensions

Dimension	F(1,23)	p	ϵ
Mental Demand	19.700	$p < 0.05$	0.10
Physical Demand	14.520	$p < 0.01$	0.07
Temporal Demand	7.720	$p < 0.05$	0.05
Performance	11.800	$p < 0.01$	0.10
Effort	48.600	$p < 0.001$	0.22
Frustration	15.120	$p < 0.01$	0.07
Workload	35.254	$p < 0.01$	0.20

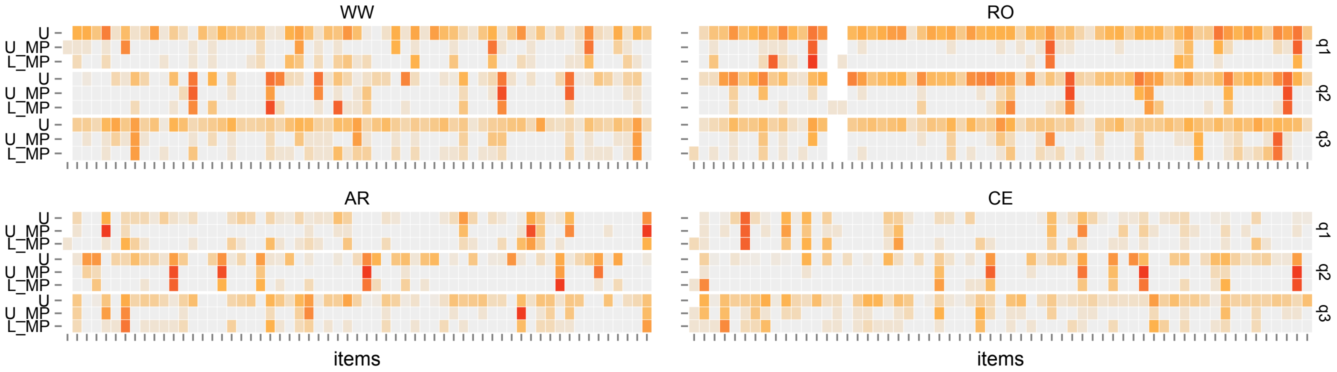


Figure 2: Correlation heatmap. Most popular items collected with our tool (U_MP) had high scores in the topic based ranking (U). Most popular items collected with the list (L_MP) are more widespread. The ranking (U), produced many high scoring items (WW-q1, RO-q1, RO-q2), indicating that personalized ranking may be more appropriate.

Table 2: Correlation analysis: ICCs established good to excellent correlations between most-popular items collected with List(L-MP), our tool (U-MP), and the ranking scores.

Task	ww-q1	ww-q2	ww-q3	ro-q1	ro-q2	ro-q3
ICC	.658	.862	.751	.857	.835	.738
Task	ar-q1	ar-q2	ar-q3	ce-q1	ce-q2	ce-q3
ICC	.814	.869	.813	.911	.866	.707

We performed intra-class correlations (ICC), using a two-way, consistency, average measures model. Results are summarized in Table 2. For broad exploration (q1 & q2), we found good to excellent ICCs. A closer look at the distribution of scores in Fig. 2 underlines the fact that high ranked documents (U) were a popular choice with U_MP and also relatively popular with L_MP. For q3, the ranking (U) produced widespread scores with less individual favorites, items L_MP were generally least popular. U_MP resulted in the most focused of the three (less blocks with higher intensity).

4 Discussion and Outlook

Results show that the fast-ranking method in our content recommender helps users quickly reorganize collections. The preference elicitation method was well received and quickly adopted. Participants experienced less effort and overall workload using our tool. Still, they took time to check their choices carefully in both U and L conditions. Comparing most popular choices after the experiment reinforces our assumptions: the fast-ranking method (U) correlates with most popular choices made with the tool (U_MP) but also without it (L_MP). Yet, widespread results in some cases call for a personalized recommendation method. Our preference elicitation forms the backbone of personalized recommendations. In the future we will explore recommendations of related items in context, showing keywords used to collect the item, other items collected together and under which collections.

Acknowledgments

This work is partially funded by CONICET (project Visual-I-Lab, Res. 4050-13) and by Know-Center. Know-Center is funded by the Austrian Research Promotion Agency (FFG) under the COMET Program.

References

[Blei, 2012] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April 2012.

[Cremonesi *et al.*, 2012] Paolo Cremonesi, Franca Garzottto, and Roberto Turrin. User effort vs. accuracy in rating-based elicitation. In *Proc. of the Sixth ACM Conf. on Recommender Systems*, RecSys '12, pg. 27–34, New York, NY, USA, 2012. ACM.

[Kay, 2006] Judy Kay. Scrutable adaptation: Because we can and must. In Vincent P. Wade, Helen Ashman, and Barry Smyth, editors, *AH*, volume 4018 of *Lecture Notes in Computer Science*, pg. 11–19. Springer, 2006.

[Knijnenburg *et al.*, 2012] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, October 2012.

[Pu *et al.*, 2011] Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. In *Proc. of the Fifth ACM Conf. on Recommender Systems*, RecSys '11, pg. 157–164, New York, NY, USA, 2011. ACM.

[Ricci *et al.*, 2011] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. *Introduction to Recommender Systems Handbook*, pg. 1–35. Springer US, 2011.

[Swearingen and Sinha, 2001] K. Swearingen and R. Sinha. Beyond algorithms: An hci perspective on recommender systems. In *ACM SIGIR. Workshop on Recommender Systems*, volume Vol. 13, Numbers 5-6, pg. 393–408, 2001.

[Tintarev and Masthoff, 2012] Nava Tintarev and Judith Masthoff. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):399–439, October 2012.