Burst the Filter Bubble: Using Semantic Web to Enable Serendipity

Valentina Maccatrozzo

The Network Institute
Department of Computer Science
VU University Amsterdam, The Netherlands
v.maccatrozzo@vu.nl

Abstract. Personalization techniques aim at helping people dealing with the ever growing amount of information by filtering it according to their interests. However, to avoid the information overload, such techniques often create an over-personalization effect, *i.e.* users are exposed *only* to the content systems assume they would like. To break this "personalization bubble" we introduce the notion of *serendipity* as a performance measure for recommendation algorithms. For this, we first identify aspects from the user perspective, which can determine level and type of serendipity desired by users. Then, we propose a user model that can facilitate such user requirements, and enables serendipitous recommendations. The use case for this work focuses on TV recommender systems, however the ultimate goal is to explore the transferability of this method to different domains. This paper covers the work done in the first eight months of research and describes the plan for the entire PhD trajectory.

1 Research Problem

We are living the Information Age - previously unfindable or unreachable information is accessible instantly and the amount of it is constantly growing. Through personalization techniques we often get to see only the chunk that relates to our interests, preventing us from being overwhelmed. Various information providers typically gather user behavior and interests data to provide personalized recommendations, *e.g.* Amazon[1], Netflix[2]. However, such information filters have downsides too. On one hand, users are constantly missing something without noticing it, and, on the other, they are getting continuously the same type of recommendations. In 2011 Pariser [18] coined a new concept to describe this phenomenon: *the filter bubble*, *i.e.* personalization filters are building around us invisible barriers that keep away the content that does not fit completely with our profiles.

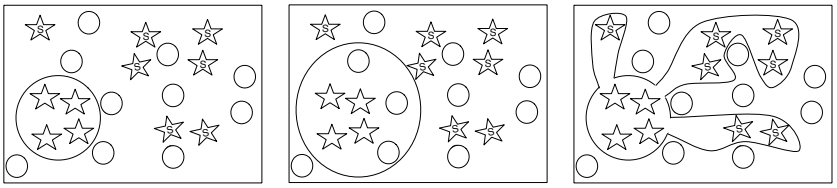
Think when you buy a book in a bookstore. You browse around the shelves letting titles and covers attract your attention. How many times it happens

[1] http://www.amazon.com
[2] http://www.netflix.com

P. Cudré-Mauroux et al. (Eds.): ISWC 2012, Part II, LNCS 7650, pp. 391–398, 2012.

you found an interesting book on a shelve you look at only by chance? This is an *unexpected encounter*. The ability to make fortunate discoveries by accident is called *serendipity*. The word was coined by Horace Walpole in a letter he exchanged with Horace Mann in 1754 [24]. He describes serendipity as "*[...] making discoveries, by accidents and sagacity, of things which they were not in quest for [...]*". Personalization as we know it in the online bookstores does not allow this to happen anymore. It makes it difficult to discover what we did not know we were looking for.

This over-personalization problem cannot be solved by simply relaxing the filters, *i.e.* by keeping the bubble bigger, because of two reasons (see Fig. 1). First, browsing through irrelevant results in an online system is not as pleasurable as browsing through a physical store - the amount is way too big, and a bird-eye view is usually not available. Second, such approach does not account explicitly for the serendipity effect, *i.e.* as serendipity is subjective, the user model has to be able to surface items that are relevant but enough novel and diverse from the standard user interests.



(a) (b) (c)

Fig. 1. This example shows how over-personalization harms recommendations. In (1a) the circle indicates an over-personalized recommendation, which includes *only* core relevant items, (1b) shows a relaxed personalized recommendation, which contains *many* irrelevant items and (1c) shows the target recommendation, which contains *all* relevant items including the serendipitous ones.

While traditional personalization approaches focus mainly on getting results as close as possible to the user profile and do not account explicitly for the serendipity effect, more recently, a trend to focus more on approaches to get serendipitous results in recommendations [1,16,25] has developed.

2 Research Context

The focus of this work is on recommender systems in the TV domain (not only movies, but more about TV programs, *e.g.* talk shows, live shows, particular episodes of series). As described above, serendipity in the context of recommender systems is represented by a well balanced mix of diversity, novelty and relevance of the recommended items with respect to the users' interests. Thus, it can be measured only with respect to a given user profile, and the challenge

is how to determine the ideal distance from the user profile in a given context, that will be still relevant. Our proposal is to use Semantic Web techniques, in particular Linked Open Data (LOD), as a means to induce novel and relevant concepts in the user profile and thus explicitly support serendipity in recommender systems.

The rich link structure and the uniform representation make the LOD cloud a good candidate to explore for 'deep' and 'novel' connections between concepts. The LOD cloud can be seen as a structured knowledge space covering a multitude of different domains (many relevant to TV, *e.g.* music, books, movies, art), where each node in the graph is a separate knowledge element and the mechanism for discovery can be applied by creating bindings between different elements.

The goal of this research is to define and develop a method for an interactive recommendation approach, where the central novelty is the discovery and utilization of serendipitous bindings between the user profile and elements previously unlinked to it. We refer to such bindings as content patterns [19] - well connected concepts in one or across LOD datasets.

The requirements, as well as initial experiments with LOD patterns for this research have been gathered and performed during the NoTube project[3]. The next stage of this research will be performed in the context of the ViSTA-TV[4] European project. The consumers anonymized viewing behavior as well as the actual video streams from broadcasters and IPTV transmitters provided by ViSTA-TV will be used as training and test data for this PhD research. The ultimate goal is to integrate the serendipity-aware recommendation strategies together with a holistic live-stream data mining analysis in a personalized electronic program guide.

3 Research Questions

The central concept of this PhD research is serendipity and its utilization in serendipity-aware recommendation algorithms. Serendipity is typically an implicit user-subjective notion that is difficult to capture in objective terms. In order to realize it in a general recommendation approach we need to identify its objective characteristics in different user contexts, and define an explicit method to measure it. This guides our first research question:

> *Can we define a method to measure serendipity for individual content elements, as well as for the overall result of a recommendation system considering an explicit user profile? Which elements of this method are domain dependent and which could be generalized?*

As the serendipity level and its success should be assessed from a user perspective we use results from previous work, in the TV and cultural heritage domains, on identifying user needs and understanding of 'serendipity' as initial

[3] `http://www.notube.tv`
[4] `http://www.vista-tv.eu`

input for this research question. We have also explored several LOD sources, discovered relevant content patterns and analyzed their statistics as possible input for the serendipity measure. Further, user surveys in the context of the concrete ViSTA-TV use cases will be performed to gather additional requirements for the definition of serendipity and for the model to assess it. A number of experiments with the serendipity-aware recommendations will be needed in order to identify the optimal serendipity level in the different use cases. Finally, similar experiments will be performed in a different domain to investigate the cross-validity of our model.

The second challenge in this work relates to the use of LOD as a structured knowledge space to discover content patterns suitable to surface serendipitous recommendations. Considering the size of the LOD cloud and the diversity of domains, types of relationships and concepts it covers, keeping the right level of relevance in the recommendation results could be a tedious task. One way of addressing this issue could be through maintaining an up-to-date user context. Therefore, the second research question is:

> *Can we use social networks activities to form a continuously evolving and relevant context of the user interests? Can we map these user interests to LOD concepts in order to discover novel user interests through LOD content patterns?*

Results from previous and related research on social activities as input for a user profile were studied. An initial set of requirements for the user profile were derived. This set should be extended and finalized through experiments in the ViSTA-TV use cases, *i.e.* applying LOD browsing procedures (Section 4) guided by a user model. The NoTube mapping of LOD concepts to user interests is used as a baseline and further extended. Experiments will be performed to determine the impact of alternative user models and their LOD mappings on the serendipity level of the recommendation results and the user satisfaction.

What is serendipitous today, may not be true tomorrow, as it is with most of the user interests. In order to be sustainable over time, recommendation strategies need to account for the decay in user interests and changes in user context that determine the serendipity aspects. So, the third question is:

> *How does the time affect the serendipity function of a recommender system? What user feedback can help to determine a possible decay in the user interests?*

In order to measure the influence of time we need to perform long-term user tests monitoring the evolution of individual user interests, the context switching and the corresponding user feedback in the whole process. We envision comparing user profile states in different moments of time and applying a set of content patterns to analyze the differences in the serendipity perception.

In the next section we are discussing the overall approach to answering the research questions and implementing the solutions.

4 Approach

Our approach combines technologies from two fields, *i.e.* user modeling and semantic-based recommendation systems. According to André et al. [4], to induce serendipity we need a common language model, so that barriers between different fields can be removed and novel connections can be established. In other words, we need to express all the components involved in the same way. Thus, centrally to this approach is the *enrichment of our data with LOD concepts*. This includes both user activities, user interests and program metadata. The enriched data enables the alignment of concepts between the user profile and the program descriptions, and subsequently the querying for related users and programs, for example through analogy, metaphors, synonymy, homonymy, and hierarchy. Here we reuse existing metadata enrichment experiences in other domains, such as in cultural heritage for defining semantic search paths from experts behavior [14], for enriching museum metadata with historical events [23], and for recommendation-based browsing through museum collections [5]. In this project we use Web services, such as Lupedia[5], to realize the enrichment of the program metadata and the user activities.

The next major step in the approach is to *find the interesting paths* in these graphs (*i.e.* content patterns) that would lead to serendipitous recommendations. We identify three such 'routes' to serendipity, *i.e.* (1) variation & selection, (2) diverging & converging and (3) analogy.

Variation & Selection. According to Campbell [8], a combination of blind variation and selective retention of concepts is the process at the basis of creative thinking. We can apply this rationale to the querying of LOD sources by deriving new concepts from the ones that are present in the user-profile and then select those that are potentially serendipitous. The selection process needs to be trained by the feedback of the user, so that the serendipitous variations can be identified. In terms of content patterns: we select new concepts following a specific pattern, and if the feedback is positive we keep on applying it, otherwise we eliminate it. Once we have a list of serendipitous patterns, *i.e.* patterns that lead to serendipitous concepts, the identification of new ones is performed on the basis of their characteristics (*e.g.* same length, same predicates but different order).

Diverging & Converging. According to Guilford [13], divergent thinking is the capacity to consider different and original solutions to one problem and is the main component in the creativity process. Convergent thinking, instead, is the ability of bringing all the solutions together and elaborate a single one. Analogously, in querying the LOD we can first discover all possible paths starting from one node in the user profile (diverging phase). Then we can identify a new node that connects all (or the most of) these new concepts together (converging phase), and use it as a serendipitous candidate.

Analogy. According to Gentner [10], an analogy is a mapping of knowledge from one domain (the base) into another (the target). In other words, a system

[5] `http://lupedia.ontotext.com`

of relations that holds among the base objects also holds for the target objects. This process, called analogical mapping, is a combination of matching existing predicate structures and importing new predicates. Following the same reasoning, we can derive analogues LOD patterns using nodes (starting from the user profile) that share (the same or similar) predicates and exchange their predicates to define new connections.

5 Related Work

Serendipity has been recognized as an important component in many fields, such as scientific research [9], art [22] and humanistic research [20]. The main point of study, especially in creative thinking, has been the strive for understanding how different serendipitous encounters take place [7].

The role of serendipity in recommender systems has also been studied. Abbasi et al. [1] examine the over-specialization problem in recommenders. Similarly to our approach, they propose a system were items are grouped in regions and recommendations are built taking items also from regions under-exposed to the user. However, contrary to our approach, they do not exploit content semantics. Oku and Hattori [16] introduce serendipity in recommendations by selecting new items mixing the features of two user-input items. This approach measures serendipity only considering past activities of the users. This differs from our approach, that does not aim necessarily at improving accuracy with respect to other recommendation techniques, but improving the overall user experience. Zhang at al. [25] present a music recommender that combines diversity, novelty and serendipity of recommendation at a slightly cost of the accuracy.

On the side of semantic recommenders, Oufaida and Nouali [17] propose a multi-view recommendation engine that integrates collaborative filtering with social and semantic recommendation. They build users' profiles and neighborhoods with three dimensions: collaborative, socio-demographic and semantic. They show how semantics enhance precision and recall of collaborative filtering recommendations. However our approach aligns more with the work done in the CHIP project[6] on a content-based semantic art recommender, where [5] explores a number of semantic relationships and patterns that allow for introducing surprisingly interesting results. One of the aim addressed by researchers in the field of semantic recommender systems is the reliability and precision of the recommended items. To tackle this issue trust network have been used. For instance, Ziegler [26] proposes suitable trust metrics to build trust neighborhoods, and to make collaborative filtering approaches applicable to decentralized architecture, *i.e.* the Semantic Web. Golbeck and Hendler [12] propose a collaborative recommender system for movies, using FOAF [6] vocabulary as a base to build a social network of trust. An example of a semantic recommender for multimedia content is given by Albanese et al. [3] that computes customized recommendations using semantic contents and low-level features of multimedia objects, past behavior of individual users and behavior of the users community as a whole. The effectiveness of the approach is evaluated on the basis of user satisfaction.

[6] `http://www.chip-project.org`

Semantic user models to enhance personalized semantic search have been researched by Jiang and Tan [15]. They propose a user ontology model that utilizes concepts, taxonomic and non-taxonomic relations in a given domain ontology to capture the users interests. Ghosh and Dekhil [11], on the other hand, discuss accurate models of user profiles using Semantic Web technologies, by aggregating and sharing distributed fragments of user profile information spread over multiple services. Related to our proposal are also the semantic user modeling from social network. Abel et al. [2] introduce a framework for user modeling on Twitter which enriches the semantics of Twitter messages and identifies topics and entities mentioned in them and, similarly to van Aart at al. [21], shows how semantic enrichment enhances the variety and the quality of the generated user profiles.

6 Future Work and Conclusions

This PhD research is now approaching the second year. Current work involves analyzing specific techniques to select possible serendipitous patterns from different LOD datasets, namely *LinkedMDB*[7] and *DBpedia*[8]. We are also investigating different techniques of enrichment, exploring natural language processing methods. The plan for the near future is to start the users surveys to gather preliminary data about their serendipity perception. Afterwards, we will follow the steps presented in Section 4.

Acknowledgments. This research is supported by the FP7 STREP "ViSTA-TV" project, as well as partially supported by the FP7 IP "NoTube" project and the ONR Global NICOP "COMBINE" project.

References

1. Abbassi, Z., Amer-Yahia, S., Lakshmanan, L.V.S., Vassilvitskii, S., Yu, C.: Getting Recommender Systems to Think Outside the Box. In: RecSys 2009, pp. 285–288 (2009)
2. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Analyzing User Modeling on Twitter for Personalized News Recommendations. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) UMAP 2011. LNCS, vol. 6787, pp. 1–12. Springer, Heidelberg (2011)
3. Albanese, M., d'Acierno, A., Moscato, V., Persia, F., Picariello, A.: A Multimedia Semantic Recommender System for Cultural Heritage Applications. In: ICSC 2011, pp. 403–410 (2011)
4. André, P., schraefel, mc., Dumais Teevan, S.T.: Discovery Is Never by Chance: Designing for (Un)Serendipity. In: C & C 2009, pp. 305–314 (2009)
5. Aroyo, L., Stash, N., Wang, Y., Gorgels, P., Rutledge, L.: CHIP Demonstrator: Semantics-Driven Recommendations and Museum Tour Generation. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ISWC/ASWC 2007. LNCS, vol. 4825, pp. 879–886. Springer, Heidelberg (2007)

[7] `http://www.linkedmdb.com/`
[8] `http://www.dbpedia.org/`

6. Brickley, D., Miller, L.: FOAF Vocabulary Specification 0.97. Namespace document, W3C (January 2010)
7. Chaomei, C.: Turning Points. The Nature of Creative Thinking. Springer (2011)
8. Campbell, D.T.: Blind Variation and Selective Retention in Creative Thought as in Other Knowledge Processes. Psychological Review 67, 380–400 (1960)
9. Garcia, P.: Discovery by Serendipity: a new context for an old riddle. Foundations of Chemistry 11, 33–42 (2009)
10. Gentner, D.: The mechanisms of analogical learning. In: Vosniadou, S., Ortony, A. (eds.) Similarity and Analogical Reasoning, pp. 199–241. Cambridge University Press (1989)
11. Ghosh, R., Dekhil, M.: Mashups for semantic user profiles. In: WWW 2008, pp. 1229–1230 (2008)
12. Golbeck, J., Hendler, J.: FilmTrust: movie recommendations using trust in web-based social networks. In: CCNC 2006, pp. 282–286 (2006)
13. Guilford, J.P.: The Nature of Human Intelligence. McGraw-Hill, New York (1967)
14. Hildebrand, M., van Ossenbruggen, J.R., Hardman, H.L., Wielemaker, J., Schreiber, G.: Searching In Semantically Rich Linked Data: A Case Study In Cultural Heritage. Technical Report INS-1001, CWI (2010)
15. Jiang, X., Tan, A.: Learning and inferencing in user ontology for personalized Semantic Web search. Information Sciences 179(16), 2794–2808 (2009)
16. Oku, K., Hattori, F.: Fusion-based Recommender System for Improving Serendipity. In: DiveRS 2011, pp. 19–26 (2011)
17. Oufaida, H., Nouali, O.: Exploiting Semantic Web Technologies for Recommender Systems: A Multi View Recommendation Engine. In: ITWP 2009 (2009)
18. Pariser, E.: The Filter Bubble. What the Internet is hiding from you. Penguin Press HC (2011)
19. Presutti, V., Aroyo, L., Adamou, A., Schopman, B., Gangemi, A., Schreiber, G.: Extracting Core Knowledge from Linked Data. In: COLD 2011 (2011)
20. Quan-Haase, A., Martin, K.: Digital Humanities: the continuing role of serendipity in historical research. In: iConference 2012, pp. 456–458 (2012)
21. van Aart, C., Aroyo, L., Brickley, D., Buser, V., Miller, L., Minno, M., Mostarda, M., Palmisano, D., Raimond, Y., Schreiber, G., Siebes, R.: The NoTube Beancounter: Aggregating User Data for Television Programme Recommendation. In: SDoW 2009 (2009)
22. van Andel, P.: Anatomy of the Unsought Finding. Serendipity: Origin, History, Domains, Traditions, Appearances, Patterns and Programmability. The British Journal for the Philosophy of Science 45(2), 631–648 (1994)
23. van Erp, M., Oomen, J., Segers, R., van de Akker, C., Aroyo, L., Jacobs, G., Legêne, S., van der Meij, L., van Ossenbruggen, J.R., Schreiber, G.: Automatic Heritage Metadata Enrichment With Historic Events. In: MW 2011 (2011)
24. Walpole, H.: To Mann, Monday 18 January 1754. In: Lewis, W.S. (ed.) Horace Walpole's Correspondence, vol. 20, pp. 407–411. Yale University Press (1960)
25. Zhang, Y.C., Séaghdha, D., Quercia, D., Jambor, T.: Auralist: Introducing Serendipity into Music Recommendation. In: WSDM 2012, pp. 13–22 (2012)
26. Ziegler, C.-N.: Semantic Web Recommender Systems. In: Lindner, W., Fischer, F., Türker, C., Tzitzikas, Y., Vakali, A.I. (eds.) EDBT 2004. LNCS, vol. 3268, pp. 78–89. Springer, Heidelberg (2004)