Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2019) 27: 2379 – 2394
© TÜBİTAK
doi:10.3906/elk-1812-25

Web personalization issues in big data and Semantic Web: challenges and opportunities

Bujar RAUFI[*], **Florije ISMAILI**, **Jaumin AJDARI**, **Xhemal ZENUNI**
Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, North Macedonia

Received: 04.12.2018 • **Accepted/Published Online:** 05.05.2019 • **Final Version:** 26.07.2019

Abstract: Web personalization is a process that utilizes a set of methods, techniques, and actions for adapting the linking structure of an information space or its content or both to user interaction preferences. The aim of personalization is to enhance the user experience by retrieving relevant resources and presenting them in a meaningful fashion. The advent of big data introduced new challenges that locate user modeling and personalization community in a new research setting. In this paper, we introduce the research challenges related to Web personalization analyzed in the context of big data and the Semantic Web. This paper also introduces some models and approaches that can bridge the gap between the two. Future challenges and opportunities related to Web personalization, analyzed from the big data and Semantic Web perspective, are also presented. The research challenges outlined in this paper involve the scrutability of user models in personalization, generic personalization, meta-personalization, open corpus personalization, and semantic data modeling.

Key words: Web personalization, user modeling, adaptive techniques, big data, Semantic Web

1. Introduction

The proliferation of data has rendered the Web space a unique environment where content is not simply searched and read but also explored and presented in a meaningful and user-friendly form. From various Web applications to social media and from mobile devices to wearables and the Internet of Things (IoT), all these innovations have together contributed to altering the Web as a global data space in their own manner. These alterations shape user interaction in a manner that profoundly transforms the human experience in the digital era. The research field of Web personalization has been very vibrant and active in the past twenty years, resulting in a plethora of terms, methods, techniques, and technologies [1]. The concept was first introduced by Brusilovsky in 1996 in the context of adaptive hypermedia systems (AHSs). A few revisions were introduced in 1998 [2], and in 2009 Knutov [3] completely revised it.

Web personalization is considered a process that consists of building models of individual user goals, preferences, and knowledge, as well as using such models throughout each interaction with users to adapt the proffered content to their preferences. User models (UMs) are utilized for automating the delivery of personalized content. They are generated by observing the knowledge gained from one topic or concept and abstracted to other, potentially wider, topics and concepts of similar interest and their continuous update and evolution.

In the last two decades, many approaches have been developed for systems that deliver Web personalization. Although it is almost impossible to provide an exhaustive list, we mention a few from the point of view of models and systems. The models we can name are the tower model [4] with the adaptive hypermedia appli-

[*]Correspondence: b.raufi@seeu.edu.mk

cation model (AHAM) [5], the Munich model [6], the Goldsmith model with Goldsmith's adaptive hypermedia application model (GAHM) [7], and layered models [8, 9], including the layered adaptive hypermedia system authoring model (LAOS). The developed systems that we can name that substantially influenced research in the adaptive hypermedia realm are AHA! [10], KBS Hyperbook [11], APeLS [12], ELM-ART [13], and Interbook [14]. It is also worth mentioning some recent solid developments, such as TANGOW and TANGOW-based systems [15], GOMAWE [16], and CoMoLe [17].

The advent of big data and the Semantic Web with linked open data technologies introduces new contexts and brings additional research challenges and opportunities to Web personalization. The vastness, robustness, and versatility of data move Web personalization into a completely new research setting. In the following subsections, several research challenges are analyzed and we offer some potential solutions within the contexts mentioned above.

1.1. Big data and web personalization

Data proliferation and the advent of big data as a concept have given rise to the constraint of the so-called five Vs [18]. Consequently, by current standards data are constrained by the speed at which new data are being generated, collected, and analyzed at any given time (i.e. velocity). The data are also constrained and affected by the amount of data produced every second across all online channels, including transaction records, network streams, experimental outputs, social media data, demographic records, citation data, clickstreams, log data, weather data, surveillance data, and sensor data (i.e. volume). The challenge that arises in this case is that volume cannot be managed if it cannot be quantified. The aim of this work is to offer directions for effective management of volume in big data to achieve effective Web personalization. Data are also affected by their structural diversity, which imposes the limit that no single data model can capture all the data's elements and features (i.e. variety). Furthermore, data variety also influences data such that they are changed quickly, which results in inconsistency (i.e. variability), and finally, a huge amount of data always negatively affects data provenance, trustworthiness, and quality. Data vastness calls data quality into question, as seen from the perspective of reliable sources, and as a result, data quality should be valued over quantity (i.e. veracity).

Based on the above, the first research challenge can be formulated as follows.

Research Challenge 1: *How can a huge amount of data be explored to achieve meaningful and actionable knowledge for effective Web personalization?*

A myriad of techniques exist, varying from pattern identification to predictive modeling, and big data analytics methods can be separated into model-driven and pattern-driven methods. Techniques such as opinion mining, record linkage, security analytics, classification and clustering, and pattern mining tend to drive this separation. With these aspects in mind, in the following subsection we outline two possible Web personalization models based on pattern- and model-driven analytics, respectively.

1.1.1. Web personalization with pattern-driven analytics

Pattern-driven analytics represents the discovery and visualization of recurring patterns in large datasets. It is mostly a quantitative approach consisting of two main paradigms: pattern discovery through sampling and aggregation and pattern discovery through thresholding and filtering.

Sampling and aggregation pattern discovery is a query-based pattern aggregation, where data can be adaptively presented based on some initial ideas (hypotheses) of the subject that is being searched. The user initially introduces a topic in which that s/he is interested and, based on that, a query is formulated and executed against the dataset. From the query, patterns are retrieved and aggregated based on some heuristic criteria. Many machine learning approaches have recently been used for pattern generation and aggregation, such as schema and documentation summarization in relational databases [19], the faceted approach [20], and spectral clustering techniques for summarization [21]. Figure 1 illustrates a model proposed for personalized content delivery using sampling and aggregation pattern discovery. Thresholding and filtering pattern discovery is based on the process of sifting through the entire dataset to search "interesting" patterns without the context of a query. The process starts with some initial "interestingness" criteria, from which patterns are extracted and later filtered and appropriately segmented. A proposed model for Web personalization using thresholding and filtering pattern discovery methods is illustrated in Figure 2. The disadvantage of the approaches mentioned above lies in the fact that they are domain-specific and do not holistically address the vastness of the data in hand. An additional problem facing both approaches for Web personalization mentioned above is that of accessing the system for the first time where no previous user data are available. This is also known as "cold start" in user modeling. The provision of initial information to handle cold start cases, where data are characterized by a high degree of velocity, volume, variety, variability, and veracity, is a research challenge for the future. There are some recent approaches that tend to address the cold start problem in recommender systems by combining classification, semantic, and heuristic techniques [22]. However, these approaches remain untested on the big data scale.

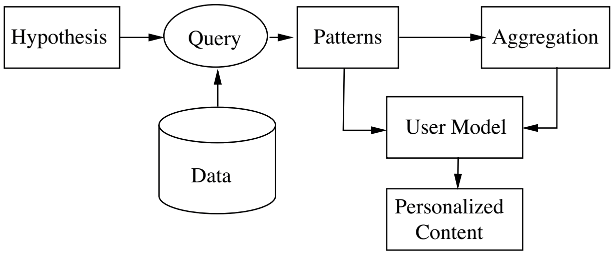

Figure 1. Personalization model using sampling and aggregation pattern discovery.

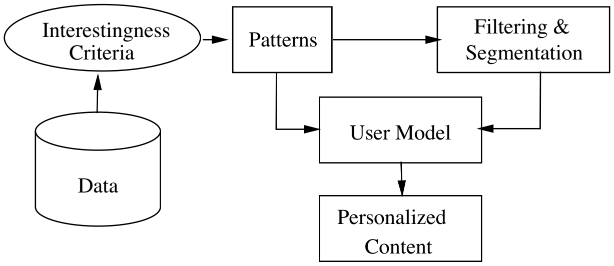

Figure 2. Personalization model using thresholding and filtering pattern discovery.

Consequently, the second research challenge can be formulated as follows.

Research Challenge 2: *Can we optimize the cold start problem in Web personalization operating on big data? Do heuristics exist for providing initial data for user modeling?*

One possible approach, together with the steps involved for providing initial data, is summarized as follows.

1. The system initiates a random number of data sources and assigns a random relevance probability.

2. The user initiates an interaction with the system.

3. New sources are retrieved based on information relevance, and probabilities are updated.

4. If further clicks and searches are initiated, Steps 2 and 3 are repeated.

5. Data source relevance and probabilities are updated in every step between Steps 2 and 4.

An example of the above steps is summarized in the pseudocode given below. Lines 2–4 initialize the sources with random probabilities and lines 6–11 check for user clicks and calculate similarity measures between the sources clicked and random sources already in the repository. On each click, this repository is updated with increasingly relevant sources.

```
1:  procedure Cold Start
2:  initialization:
3:      $R[n] \leftarrow$ array of data source
4:      $random\_probablities \leftarrow R[initial\ probabilities]$
5:  loop:
6:      while $user\_click$ do
7:          if $user\_click = true$ then return relevant sources related to the click
8:              $probablities \leftarrow similarity\_measure(R[i])$
9:  loop $similarity\_measure$:
10:         for $j \leftarrow 1$ to $length(R[n])$ do
11:             $R[j] \leftarrow similarity(R[i], R[j])$
12:     goto loop.
```

This model represents a randomized probabilistic algorithm, where there exist cost c_i associated with every resource R and also cost c_s for selecting that resource. Based on this, an informal algorithm complexity for the worst case yields a complexity of $O(c_i n + c_s m)$, where n is the total number of resources and m is the number of times a best resource is selected. If we consider an average case scenario, we are interested in the number of times we select a suitable resource. For this, let X_i be the indicator random variable that indicates whether a resource i is selected, i.e. let $X_i = 1$ if resource i is selected, and 0 otherwise.

Let $X = \sum_{i=1}^{n} X_i$ be the number of times a new resource is selected. We are interested in finding a resource $R[X_i]$, which, by a property of expectation from line 3, is just $\sum_{i=1}^{n} R[X_i]$. From the pseudocode, we need to find the probability of $R[X_i] = P$ and to do this we assume that resources are selected in a random order (this is ensured by the randomization step enforced in line 4). Any one of the resources is equally likely to be the best qualified thus far, and this can be represented as $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + ... + \frac{1}{n}$. Therefore, the probability that resource i is better than resources 1 to $i-1$ is just $1/i$. Therefore, $R[x_i] = 1/i$, and $R[X] \sum_{i=1}^{n} 1/i = \ln n + O(1)$. Finally, the expected number of selected resources is $O(\ln n)$ and the cost is $O(c_s \ln n)$.

1.1.2. Web personalization model with model-driven analytics

Model-driven analytics tends to generalize the current knowledge at hand with other knowledge that is related to it. The generalized knowledge in this case is a latent concept that is discovered. At its core, model-driven analytics is a process of pattern discovery coupled with semantic modeling through ontology engineering. Model-driven analytics is usually viewed through the prism of descriptive models and predictive models. A descriptive model attempts mainly to fit a model to describe the observed data, while predictive models are designed to discover a model that can predict the values of data elements for the future. The discovery of latent concepts is observed in many closed corpus domains [23]. The goal is to demonstrate the manner in which the inclusion of Semantic Web technologies can be deployed to enrich semantic vocabularies for the purpose of providing semantic browsing, searching, and semantic personalization. After the semantic enrichment of data is complete, the process of discovering latent concepts in such a closed corpus is accomplished. For example, let us consider the process of latent concept discovery in an art museum, consisting of the following steps.

1. A user clicks on a specified artwork or topic.
2. Based on that specific artwork or topic, other resources, such as rated artists and rated topics, are presented.
3. After several clicks, recommendations related to topics and artists emerge, such as the artist's other paintings and the timeline of their appearance.

A typical interaction scenario adheres to the following path. The user clicks on a specific artwork, such as *The Jewish Bride*, after visiting the painting. The system recommends its creator, i.e. *Rembrandt*, and also provides other painters and their artworks, for example, Rembrandt's master, Isaac van Swanenburg, and his painting *Spinning Wool*. The entire process results in a latent concept, such as *The Dutch Golden Age - Baroque*.

The problem of the approach mentioned above is its domain limitation. The overall UMs that are generated by the application are tied to a specific domain, in the above case the art museum. To design the UM such that it is practical on a large scale would be a considerable challenge. This is evident in repositories where the Web is used to link resources that were not previously linked, such as the Linked Open Data initiative [24]. Linked Open Data currently contains approximately 9960 datasets having 149,423,660,620 triples from 2973 datasets, out of which 192,230,648 are triples from 2838 dumps and 149,231,429,972 triples from 151 datasets via SPARQL.[1]

Consequently, the third research challenge that can be formulated is as follows.

Research Challenge 3: *How can a personalized resource delivery be created for huge data repositories such as Linked Open Data? How can a UM be built in such versatile environments?*

Our proposed method and "path of attack" for overcoming this challenge is to bridge the gap between big data and the Semantic Web. A possible general approach that we propose is to utilize a model process of discovery from the big data and feed it to the Semantic Web model and then utilize its inference capabilities for future model predictions. Figure 3 depicts the integration of the Semantic Web with big data in the context of Web personalization. To produce a specific model from the big data sources, pattern-driven analytics can be used for generating a model based on the observed data. This model can be mapped to an ontology on the basis of which an inference model can be generated that can serve as a UM for the personalization process. A proof of concept was introduced in [25]; however, it is worth mentioning that it has not been extensively tested on a large scale. The problem of adopting this approach on a large scale lies in the process of mapping the big data entities to Semantic Web ontologies. The mapping has proven to constitute quite a difficult task, especially when such ontologies number in the hundreds of thousands.

[1] Abele A, McCrae JP, Buitelaar P, Jentzsch A, Cyganiak R. Linking Open Data Cloud Diagram [Online]. Website https://lod-cloud.net/ [accessed 24 February 2019].

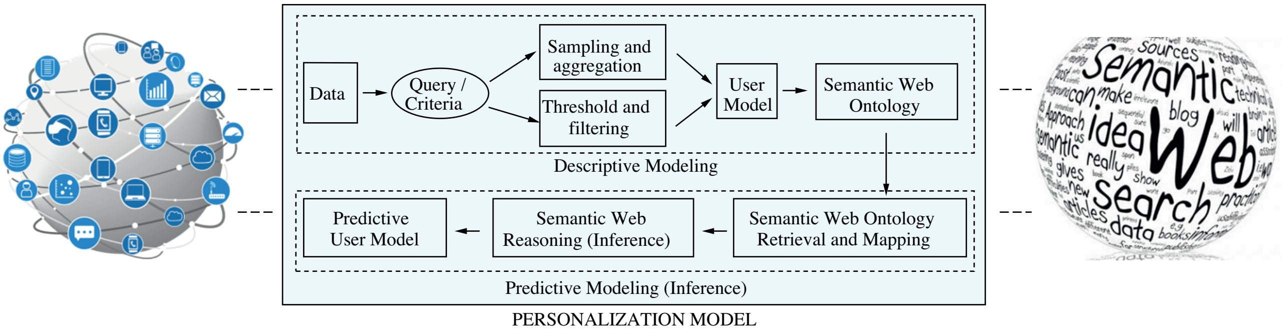

Figure 3. Personalization model with big data and the Semantic Web.

The rest of this paper is organized as follows. Section 2 introduces the state of the art of trends in Web personalization. Section 3 presents some challenges identified within user modeling, adaptation, and personalization (UMAP), together with some directions for tackling them, and Section 4 concludes this paper.

2. Related work

Almost a decade ago, several new emerging trends for personalization were envisioned, and effort has been invested toward a generic AHS [26]. These trends have shifted and the possibilities of some have been exhausted, whereas other aspects, such as meta-personalization, context awareness, and open-corpus personalization, slightly went “under the radar” of personalization research, without researchers giving this oversight serious thought. In this section, we introduce these trends and analyze them from the current perspective and the work accomplished thus far.

2.1. Ontologies for web personalization

The use of ontologies for Web personalization is an idea that has been propagated within the research community for more than a decade. It is publicly known that, for many personalized and adaptive applications, currently authors create not only the information space but also the semantic or concept space.

The process of combining UMs and content based on a single ontology across multiple applications has been shown to be feasible. This is seen especially in various applications, such as learning environments [27] [28], flexible and reusable user interfaces [29], and social network-based user modeling [30]. Some attempts were even focused on building a complete adaptive application based on ontology and defining rules and inferences on such rules on a semantic level [31]. Current studies have even attempted reasoning on more dynamic data, such as stream sensor data [32] or dynamic Semantic Web data [33].

The process of personalization and user modeling where multiple ontologies are involved has proven to be a nontrivial task, because it involves two main obstacles: mapping across multiple ontologies for the same concept and reasoning over ontologies. One proposed semantic personalization approach is to use both personalized information retrieval at a semantic level and resource personalization. The overall architecture of the system consists of the following main modules.

- *The user data extraction module* extracts user-browsing activities and inserts them into the user knowledge base (ontology) as object instances or data properties. User-browsing activities are stored in many formats (logs, database tables, etc.) on the browser’s side and additional intermediary tasks are required for these extractions and data mappings.

- *The user knowledge base* represents an ontology that depicts user activities on the semantic level. It usually comprises user sessions, visited content, and generated user views as part of personalization. Browsing activities from the data extraction phase are also stored in the ontology in the form of object instances or data properties.

- *Ontology retrieval and application* is a phase of ontology aggregation and fetching from various repositories based on user interaction behavior. For instance, if a user is interested in a particular topic, to "adapt" to this concrete user activity a specific ontology from publicly available repositories should be acquirable (such as Cupboard, Knoodl, Schemapedia, SchemaWeb, TONES, etc.).

- *The reasoning process* draws conclusions on instances and data properties based on the classes, class hierarchies, and various restriction properties defined explicitly in the ontologies.

The system has been tested against 1699 retrieved documents, which resulted in a total of 43,713 entries in the knowledge base. However, the approach is yet to be tested in big data environments [34].

2.2. Open corpus personalization

Modeling a knowledge base from a known data space and its mapping to a specific concept (closed corpus personalization) is easier than mapping a previously unknown data space to a relevant concept (open corpus personalization). In the former case, the mapping process can be executed by the author or a domain expert; in the latter case, the mapping of a previously unknown data space should be executed at runtime, which requires bringing the fields of Web personalization, big data, information retrieval, data mining, and the Semantic Web together.

Previous research on personalization was focused mainly on separating linking between resources and the space to which they relate and attempting to find alternative linking based on the open corpus space that the resources encounter. Most of the methods were focused around personalized navigation support for open corpus resources, providing either manual or automatic indexing of such open corpus resources [35].

In recent years, there has been a tendency to apply ontology-based personalization for open corpus resources [36], generate open corpuses for languages [37], or generate open corpuses by performing large-scale information retrieval, as well as language construct retrieval and linking [38].

Although the advent of big data has rendered open corpus personalization very difficult, the approach in Research Challenge 3 can still be applied by utilizing big data meaningfully through the enforcement of appropriate data integration. The work of Bizer et al. provided a solid ground for future approaches for good data integration. Originally, Bizer et al. presented a data integration step challenge, which involved a big data closed corpus of US patients' medical data [39]. These steps can be extrapolated also to an open corpus by following the following revised steps.

- *Define the subcorpus* in relation to the information space related to the problem that needs to be solved or query that should be answered, e.g., finding facts about the effects of global warming related to climate change in the last five years.

- *Search* the subcorpus space and find resources that map to the related problem.

- *Extract, transform, and load (ETL)* the relevant parts of the resource and present them in an appropriate format.

- *Perform entity resolution*, which should verify that the resources and the entities within the resources are unique, relevant, and comprehensive. Given that unique identification is practically and technically unfeasible, not all candidate data portions will refer to the entity in question. More challenging are data elements that describe aspects of the entity in question at different levels of abstraction and from different viewpoints and contexts.

- *Verify query answering efficiency*; that is, after the data element selection process for the entity, compute the answer efficiency using domain-specific comparisons and calculations.

It is worth mentioning that entity resolution can be effectively tackled by using Semantic Web technologies, especially using OWL disambiguation constructs, such as namespaces, equality, inequalities, property characteristics, disjointedness, and cardinalities.

2.3. Group personalization

Traditional personalization systems deliver personalization for a single user. However, the proliferation of data today requires optimizations. As a result, group personalization appears to be a feasible approach in the case of big data. Taking into account the effect of personalizations generated for other users and applying the same or similar personalization to a current user appears a solid foundation for future personalization methods and techniques within the realm of big data and the Semantic Web. In this direction, group personalization approaches exist that tackle the cold-start problem [40] and group personalization for friend recommendations on social media [41].

2.4. Information retrieval and data mining

Information retrieval and data mining techniques for personalization have been in existence for more than two decades. An extensive review of Web mining for personalization was presented in [42]. Data preprocessing approaches, such as similarity measures, sampling, and dimensionality reduction, have been used extensively in Web log mining for extracting user behavior in conjunction with the Semantic Web [43] [44], as have classification techniques, such as nearest neighbors, decision trees, ANNs, Bayesian classifiers, SVMs, and ensemble learning, as presented for example in [45], and clustering methods, such as k-means and specific cluster analysis for recommendations [46].

Whereas all these methods have yielded some substantial results for capturing their “immediate” effects on personalization, the advent of deep learning within big data can help in the future to generate more “evolutive” models, which will tend to grasp the more long-term effects. To the best of our knowledge, no method or model that outlines such an approach exists.

2.5. Meta-personalization

Previous personalization techniques applied in the Web environment are related mainly to analyzing user behavior to produce a personalization effect. This raises many problems, especially when the system becomes very efficient in presenting the content that it already knows and for which it is trained. However, the question emerges as to the action the system should take when users access “nonrecommended” content. This raises the issue of scrutinizing the personalization process (scrutability). The main question posed here is why that particular personalization effect occurred. This requires approaches that will monitor why specific personalization rules occur as such and update such rules where necessary; hence, the term meta-personalization is used.

Thus far, systems have been proposed that have tended to present some meta-personalization functionalities, such as KBS Hyperbook, LAOS, or AHA! as mentioned in Section 1. However, their functionality is limited to their closed domains and applications. Meta-personalization in the context of big data remains to be seen.

2.6. Context awareness and multimedia personalization

The proliferation of ubiquitous environments, such as the IoT, demands personalization methods and techniques that render personalization systems more decoupled and less dependent on the environments in which they operate. Recently, contributions in the area of personalized learning systems [47] [48] [49] or context-aware recommender systems [50] [51] have appeared. The generation of personalized information systems that will provide content independence at every level of the personalization process (data collection, UM generation, and personalization engine) is yet to be realized. One approach that we have proposed in this direction is to introduce more fine-grained and loosely coupled content types [52].

3. Future challenges and opportunities

Personalization methods and techniques cover various aspects, such as personalization of content, links, or both. Well-known personalization methods are adaptive presentation and adaptive navigation support with techniques such as annotation, hiding, appearance, or dimming of content or links. A complete taxonomy of methods and techniques was provided by De Bra and Brusilovsky [53]. To make these methods and techniques practical on a large scale is a substantial research challenge. Based on the literature review above, the following challenges and future research opportunities are identified.

1. Scrutability of UMs and personalization.
2. Perpetuity of personalization in multicontext environments.
3. Meta-personalization that avoids the filter bubble.
4. The problem of convergent and divergent semantics.

3.1. Scrutability of user modeling and personalization

Scrutability in user modeling has been a long-lasting and debatable issue in Web personalization [54]. It is not sufficient for the system to recommend specific personalized content to the user; the process of quantifying the amount of knowledge the system has about the user is also crucial. Consequently, the process of scrutinizing a particular adaptation rule in terms of the reason why it behaved as it did represents adaptation scrutability. The generation of a UM based on simple (expert system written) rules renders the UM self-explanatory and easily understandable [55]. For instance, supposing a closed corpus of data, such as in the case of the art museum mentioned above, generalization facts can be inferred by using rules in the following form:

$$visit(painting \quad A) \quad \text{by} \quad painter B \implies relevance(A) \leftarrow 100\% \quad \text{and} \quad relevance(B) \leftarrow \frac{100\%}{\#paintings(A)} \quad (1)$$

In this rule, the approach is fairly straightforward. The user is visiting a specific resource, in which case the relevance of resource A is increased by 100% and the relevance of closely related resource B is increased by the fraction proportional to the number of other resources concerning resource A. These rules can be easily templated in the UM. However, the behavior and effectiveness of scrutable models operating in the large scale remain to be explored. Consequently, the fourth research challenge addresses scrutability in terms of template rule generalizations.

Research Challenge 4: *Can we define a scrutable personalization model by utilizing an information schema for personalization and associate generated template rules with the connections in such a schema?*

The use of direct inferences available in ontologies can help in this direction, because evidence exists that such inferences are quite fast [56]. This is important, because many systems that use data mining, neural networks, and the recently developed deep learning techniques fail to grasp the semantic interpretation of a specific recommended resource. The issue of the scrutability of personalization becomes even more challenging when the personalization process is based on collective behavior involving many users. In this direction, the user modeling and personalization community, to the same extent as it considers the aspect of UM scrutability, should also consider the issue of the scrutability of personalization rules. This means that we should make personalization rules as scrutable as UMs.

3.2. Perpetuity of personalization in multicontext environments

In user modeling and personalization, rigorous empirical approaches have been used to test the effectiveness of one system as compared with others. If one method was deemed better than another, the evaluation process involved test subjects that proved the claim with a solid degree of certainty. In the study in [57], many individual UMs were combined to create a group recommendation. The approach applied many methods, such as plurality voting, averaging and multiplicative methods, the Bora count method, the Copeland rule, and the least misery strategy. However, to demonstrate the effectiveness of one system in multiple contexts has proven to be a problem. In practice, there can be a method that operates with a high level of accuracy that can be validated across multiple users and even user groups; however, when the context is changed the personalization accuracy and efficiency are diminished. In this sense, the best "engine" for user modeling and personalization has yet to be designed. Such an engine is not expected to be compartmentalized for every application type and domain as the situation currently stands, but the search should be focused on combinations of classes, methods, and types of applications, which, when combined with personalization, will yield satisfactory results in multiple contexts.

Semantic Web technologies should be considered an important niche in this process. Semantic annotation of the context of uses in one ontology would be a grounded track to explore; the user modeling and personalization research community has not yet been able to meet the challenge. The reasons vary from the overwhelming inundation of big data on the one hand to poor industry adoption of Semantic Web technologies on the other, despite the fact that the Semantic Web has been a well-founded W3C standard for more than two decades.

3.3. Meta-personalization and filter bubble prevention

Domain-specific personalization relies on domain experts to define UMs and personalization algorithms for such systems. For instance, in personalized learning environments, the expert designs the rules based on the learning environments (learner-centered, knowledge-centered, assessment-centered, community centered, etc.), learning styles (verbal, visual, and aural), or synchronous/asynchronous learning. The experts also define the sequence of learning, its pace, and the appropriate ordering of the material so that the personalization engine can set requirements for the topic, such as a prior reading list, set the progress on the topic, etc.

In cases where the domain tends to become larger, semantic knowledge bases are used together with template rules to generate personalization. To exacerbate the problem, in modern large and dynamic applications the human expert(s) approach becomes unfeasible. Thus, data-driven personalization is an effective means of providing personalized content to users. For this reason, collaborative filtering has been in existence for many years and utilization of the "wisdom of the crowd" to recommend content has been an acceptable approach. The caveat of the collaborative approach lies in the fact that it is focused on past user behaviors rather than future possible activities. As a result, the personalization tends to evolve by collectively taking into account only previous experiences. The system perfects itself by becoming increasingly better by using previously personalized content but fails to provide ordinary content that might be of interest to the user. However, what happens when someone wants to access nonrecommended content? The user becomes trapped in a set of personalized content, and when he or she cannot easily access nonrecommended content, it lurks in its data space called "the filter bubble". In this aspect, many e-commerce, news, and video content providers fail to address the issue. Figure 4 illustrates the filter bubble and possible "bursting" by using predictive modeling.

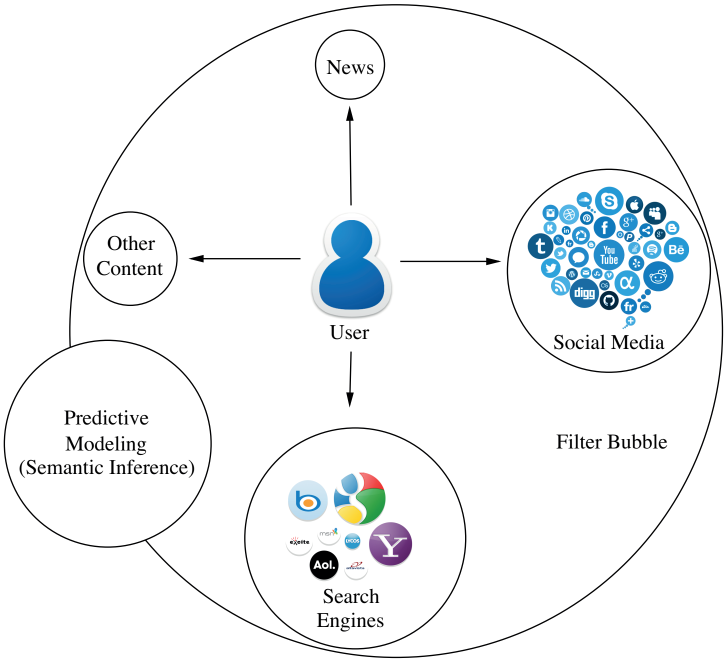

Figure 4. Predictive modeling approach for filter bubble "bursting".

One interesting research challenge that is worth exploring is as follows.

Research Challenge 5: *How can semantic data and semantic inference be utilized for meta-personalization and filter bubble avoidance?*

Semantic Web technologies provide a solid ground for inference on huge repositories. The Semantic Web languages have reached a solid maturity level in that they are capable of representing factual knowledge through data models (XML, RDFs, and OWL), terminological knowledge through ontologies (OWL, DAML+OIL, etc.), and inference knowledge (OIL, DAML+OIL, and OWL). Each of these has its own strengths and weaknesses resulting from the level of expressiveness and inference. However, the technology has reached a level where it can be used in a layered approach on a large scale [58]. Very good semantic Web knowledge bases exist, such as YAGO [59], that can be utilized for meta-personalization that could either "burst" the filter bubble for recommending new content or reduce the filter bubble by continuously feeding new nonpersonalized content in which the user might show interest.

The only caveat of the approach mentioned above is that it is not system- and method-bound. It is rather related to the structure or, more precisely, to the no-structure of the data in hand, especially in extremely large data repositories. There exist W3C initiatives, such as Linked Open Data and Microformats, that tend to help enrich the Web structure with semantic meaning for the data at hand, but their feasibility is yet to be determined.

3.4. The problem of convergent and divergent semantics

The issue of data provenance and trust in the Semantic Web was debatable from its very inception. There are many points of view related to different issues. For instance, if we consider the Middle East crisis as described in a Wikipedia article, we see that there exist different points of view on the topic: the Palestinian point of view on the Arab-Israeli conflict, the Israeli point of view, the point of view of the neighboring countries, and the point of view of the international community. If a particular semantic knowledge base was filled with facts from all points of view, these facts would be debatable on the one hand and convergent and conflicting from the personalization standpoint on the other. The above-mentioned case is a typical example of convergent semantics. Semantic knowledge today suffers heavily from convergent and divergent semantics. The convergent and divergent semantics is depicted in Figure 5. In the former case, for the same issue, there are many facts that make the trustworthiness of the source weak and a high-quality personalization process almost unattainable, whereas in the latter case a "dispersion" of facts that may contribute to the same goal of knowledge is observed. The problem that personalization faces within the context of semantic divergence is that, for the sake of quick data retrieval, inference, and content personalization, many sources fail to aggregate. If we consider counter terrorism measures, for instance, data aggregation is of utmost importance. Collaboration on building knowledge bases from various sources, such as intelligence sources, banks, local police enforcement, and social institutions, and detecting specific data patterns from such divergent environments represent a challenge on which the user modeling and personalization research community should focus.

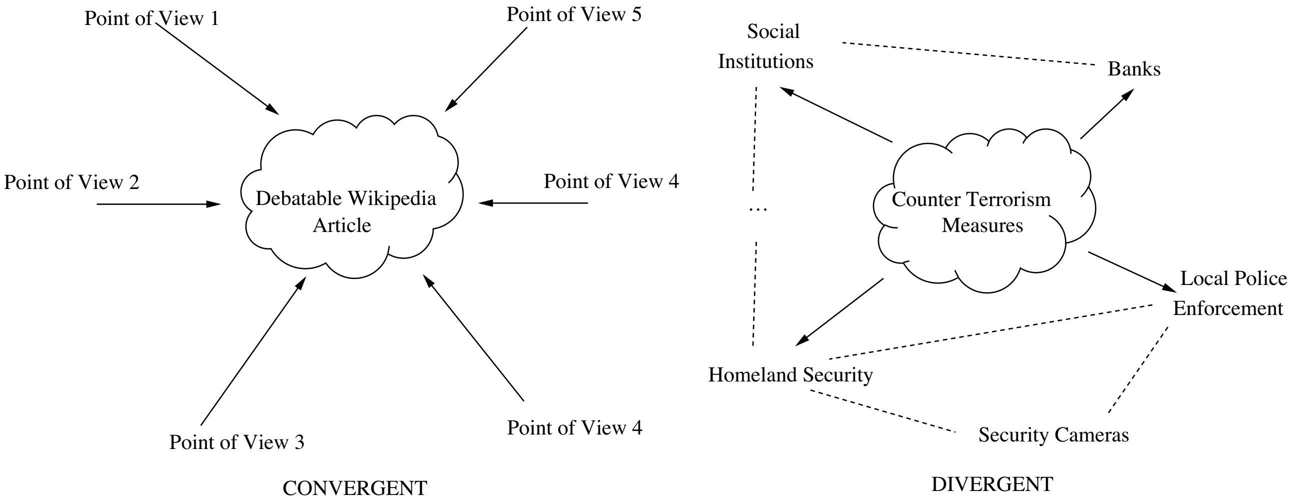

Figure 5. Convergent and divergent semantics.

4. Conclusion

The omnipresence of data will most likely shift research in the UMAP field in the coming years toward the development of methods and techniques that will deliver more generic, open-corpus, and effective applications that will meet the challenges of big data. Based on the above-mentioned identified disadvantages, the state of the art, and the research challenges lying ahead, the authors can foresee meta-personalization and personalization scrutability as directions for hot topics in UMAP in the years to come. It is also considered that the community will tackle Web personalization to bridge the gap between big data and the Semantic Web, creating systems that will incorporate both in a single holistic framework.

The purpose of this paper was threefold: first, to present a brief introduction to Web personalization from the perspective of both big data and the Semantic Web, together with the current state of the art; second, to identify potential research challenges related to the three disciplines; and third, to outline future challenges and opportunities that can attract the attention of the community.

References

[1] Brusilovsky P. Methods and techniques of adaptive hypermedia. Journal of User Modeling and User-Adapted Interaction 1996; 6 (2-3): 87-129.

[2] Brusilovsky P. Methods and techniques of adaptive hypermedia. In: Brusilovsky P, Kobsa A, Vassileva J (editors). Adaptive Hypertext and Hypermedia. Berlin, Germany: Springer, 1998. pp. 1-43. doi: 10.1007/978-94-017-0617-9_1

[3] Knutov E, Bra PD, Pechenizkiy M. AH 12 years later: a comprehensive survey of adaptive hypermedia methods and techniques. New Review of Hypermedia and Multimedia 2009; 15 (1): 5-38. doi: 10.1080/13614560902801608

[4] Bra PD, Houben GJ, Kornatzky Y. An extensible data model for hyperdocuments. In: ACM Conference on Hypertext; New York, NY, USA; 1992. pp. 222-231.

[5] Bra PD, Calvi L. AHA! An open adaptive hypermedia architecture. New Review of Hypermedia and Multimedia 1998; 4 (1): 115-139. doi: 10.1080/13614569808914698

[6] Koch N, Wirsing M. The Munich reference model for adaptive hypermedia applications. In: Adaptive Hypermedia and Adaptive Web-Based Systems Conference; London, UK; 2002. pp. 213-222.

[7] Ohene-Djan J, Fernandes A. Modelling personalisable hypermedia: The Goldsmiths model. New Review of Hypermedia and Multimedia 2002; 8 (1): 99-137. doi: 10.1080/13614560208914738

[8] Cristea A, Calvi L. The three layers of adaptation granularity. In: International Conference on User Modeling; Johnstown, PA, USA; 2003. pp. 4-14.

[9] Raufi B. Adaptive Web-Based Systems: From Conception to Implementation. Saarbrucken, Germany: Lambert Academic Publishing, 2013.

[10] Bra PD, Smits D, Stash N. The design of AHA! In: Hypertext and Hypermedia Conference; New York, NY, USA; 2006. pp. 133-134.

[11] Henze N. Adaptive hyperbooks: adaptation for project-based learning resources. PhD, University of Hannover, Hannover, Germany, 2000.

[12] Conlan O, Hockemeyer C, Wade V, Albert D. Metadata driven approaches to facilitate adaptivity in personalized eLearning systems. Journal of Information and Systems in Education 2002; 1: 38-45.

[13] Brusilovsky P, Schwarz E, Weber G. ELM-ART: An intelligent tutoring system on world wide web. In: Intelligent Tutoring Systems Conference; Montreal, Canada; 1996. pp. 261-269.

[14] Eklund J, Brusilovsky P. Interbook: an adaptive tutoring system. UniServe Science News 1999; 12 (3): 8-13.

[15] Carro RM, Pulido E, Rodriguez P. TANGOW: Task-based adaptive learner guidance on the WWW. In: Adaptive Systems and User Modeling on the Web, Second Workshop; Toronto, Canada; 1999. pp. 49-57.

[16] Balik M, Jelinek I. General architecture of adaptive and adaptable information systems. In: Enterprise Information Systems Conference; Madeira, Portugal; 2007. pp. 29-34.

[17] Martin E, Carro RM, Rodriguez P. A mechanism to support context-based adaptation in m-learning. In: First European Conference on Technology Enhanced Learning; Crete, Greece; 2006. pp. 302-315.

[18] Srinivasa S, Bhatnagar V. Big data analytics. In: Big Data Analytics BDA Conference; New Delhi, India; 2012. pp. 24-26.

[19] Yasir A, Swamy MK, Reddy PK. Exploiting schema and documentation for summarizing relational databases. In: Big Data Analytics Conference; New Delhi, India; 2012. pp. 77-90.

[20] Nambiar U, Faruquie T, Kumar S, Morstatter F, Liu H. Faceted browsing over social media. In: Big Data Analytics Conference; New Delhi, India; 2012. pp. 91-100.

[21] Gupta A, Kathuria M, Singh A, Sachdeva A, Bhati S. Analog textual entailment and spectral clustering (atesc) based summarization. In: Big Data Analytics Conference; New Delhi, India; 2012. pp. 101-110.

[22] Lika B, Kolomvatsos K, Hadjiefthymiades S. Facing the cold start problem in recommender systems. Journal of Expert Systems with Applications 2014; 41 (4): 2065-2073. doi: 10.1016/j.eswa.2013.09.005

[23] Roes I, Stash N, Wang Y, Aroyo L. A personalized walk through the museum: the chip interactive tour guide. In: Human Factors in Computing Systems Extended Abstracts (CHI'09); Boston, MA, USA; 2009. pp. 3317-3322.

[24] Bizer C, Heath T, Berners-Lee T. Linked data-the story so far. International Journal on Semantic Web and Information Systems 2009; 5: 1-22. doi: 10.4018/978-1-60960-593-3.ch008

[25] Raufi B, Ismaili F, Ajdari J, Ferati M, Zenuni X. Semantic resource adaptation based on generic ontology models. In: ICSOFT-PT'14 Software Paradigm Trends Conference; Vienna, Austria; 2014. pp. 103-108.

[26] Knutov E. Generic adaptation framework for unifying adaptive web-based systems. PhD, Eindhoven University of Technology, Eindhoven, the Netherlands, 2012.

[27] Belcadhi LC. Personalized feedback for self assessment in lifelong learning environments based on semantic web. Journal of Computers in Human Behavior 2016; 55: 562-570. doi: 10.1016/j.chb.2015.07.042

[28] Ouf S, Ellatif MA, Salama SE, Helmy YA. A proposed paradigm for smart learning environment based on semantic web. Journal of Computers in Human Behavior 2017; 72: 796-818. doi: 10.1016/j.chb.2016.08.030

[29] Khalili A, Loizou A, Harmelen FV. Adaptive linked data-driven web components: building flexible and reusable semantic web interfaces. In: International Semantic Web Conference; Kobe, Japan; 2016. pp. 677-692.

[30] Plumbaum T. User modeling in the social semantic web. PhD, TU Berlin, Berlin, Germany, 2015.

[31] Dolog P, Nejdl W. Semantic web technologies for the adaptive web. Lecture Notes in Computer Science 2007; 4321: 697-719. doi: 10.1007/978-3-540-72079-9_23

[32] Jajaga E, Ahmedi L, Ahmedi F. StreamJess: A stream reasoning framework for water quality monitoring. Journal of Metadata, Semantics and Ontologies 2016; (11) 4: 207-220. doi: 10.1504/IJMSO.2016.083507

[33] Margara A, Urbani J, Harmelen FV, Bal H. Streaming the web: reasoning over dynamic data. Journal of Web Semantics 2014; 25: 24-44. doi: 10.1016/j.websem.2014.02.001

[34] Raufi B, Ismaili F, Ajdari J, Zenuni X. Knowledgebase harvesting for user-adaptive systems through focused crawling and Semantic Web. In: CompSysTech'16 Computer Systems and Technologies Conference; Palermo, Italy; June 2016. pp. 323-330.

[35] Brusilovsky P. Adaptive navigation support for open corpus hypermedia systems. In: Adaptive Hypermedia and Adaptive Web-Based Systems Conference; Hannover, Germany; 2008; pp. 6-8

[36] Sosnovsky S, Hsiao IH, Brusilovsky P. Adaptation “in the wild”: ontology-based personalization of open-corpus learning material. In: Technology Enhanced Learning Conference; Saarbrücken, Germany; 2012. pp. 425-431.

[37] Bowman SR, Angeli G, Potts C, Manning CD. A large annotated corpus for learning natural language inference. In: Empirical Methods for Natural Language Processing Conference; Lisbon, Portugal; 2015. pp. 632-642

[38] Schmitz M, Bart R, Soderland S, Etzioni O. Open language learning for information extraction. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning; Jeju Island, Korea; 2012. pp. 523-534.

[39] Bizer C, Boncz P, Brodie ML, Erling O. The meaningful use of big data: four perspectives–four challenges. ACM Sigmod Record 2011; 40 (4): 56-60. doi: 10.1145/2094114.2094129

[40] Bobadilla J, Ortega F, Hernando A, Bernal J. A collaborative filtering approach to mitigate the new user cold start problem. Journal of Knowledge-Based Systems 2012; 26: 225-238. doi: 10.1016/j.knosys.2011.07.021

[41] Agarwal V, Bharadwaj KK. A collaborative filtering framework for friends recommendation in social networks based on interaction intensity and adaptive user similarity. Journal of Social Network Analysis and Mining 2013; 3 (3): 359-379.

[42] Eirinaki M, Vazirgiannis M. Web mining for web personalization. ACM Transactions on Internet Technology 2003; 3 (1): 1-27. doi: 10.1145/643477.643478

[43] Barla M, Tvarožek M, Bieliková M. Rule-based user characteristics acquisition from logs with semantics for personalized web-based systems. Journal of Computing and Informatics 2009; 28 (4): 399-428.

[44] Kaur N, Aggarwal H. Query based approach for referrer field analysis of log data using web mining techniques for ontology improvement. International Journal of Information Technology 2018; 10 (1): 99-110.

[45] Ghorab MR, Zhou D, O’Connor A, Wade V. Personalised information retrieval: survey and classification. Journal of User Modeling and User-Adapted Interaction 2013; 23 (4): 381-443.

[46] Zhang Y, Lai G, Zhang M, Zhang Y, Liu Y et al. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In: ACM SIGIR Conference on Research & Development in Information Retrieval; Gold Coast, Australia; 2014. pp. 83-92.

[47] Hwang GJ. Definition, framework and research issues of smart learning environments-a context-aware ubiquitous learning perspective. Smart Learning Environments 2014; 1 (4): 1-4.

[48] Sampson DG, Zervas P. Context-aware adaptive and personalized mobile learning systems. In: Sampson DG, Isaias P, Ifenthaler D, Spector JM (editors). Ubiquitous and Mobile Learning in the Digital Age. New York, NY, USA: Springer, 2013. pp. 3-17.

[49] Yang TC, Hwang GJ, Yang SJH. Development of an adaptive learning system with multiple perspectives based on students’ learning styles and cognitive styles. Journal of Educational Technology & Society 2013; 16 (4): 185-200.

[50] Adomavicius G, Tuzhilin A. Context-aware recommender systems. In: Ricci F, Rokach L, Shapira B (editors). Recommender Systems Handbook. Boston, MA, USA: Springer, 2015. pp. 191-226.

[51] Verbert K, Manouselis N, Ochoa X, Wolpers M, Drachsler H et al. Context-aware recommender systems for learning: a survey and future challenges. IEEE Transactions on Learning Technologies 2012; 5 (4): 318-335.

[52] Raufi B, Ismaili F, Zenuni X. Modeling a complete ontology for adaptive web based systems using a top-down five layer framework. In: Information Technology Interfaces Conference; Dubrovnik, Croatia; 2009. pp. 511-518.

[53] Bra PD, Brusilovsky P, Houben GJ. Adaptive hypermedia: from systems to framework. ACM Computing Surveys 1999; 31 (4): 12. doi: 10.1145/345966.345996

[54] Kay J. Stereotypes, student models and scrutability. In: Intelligent Tutoring Systems Conference; Montreal, Canada; 2000. pp. 19–50.

[55] Bra PD. Challenges in user modeling and personalization. IEEE Intelligent Systems 2017; 32 (5): 76-80.

[56] Matentzoglu N, Vigo M, Jay C, Stevens R. Inference inspector: improving the verification of ontology authoring actions. Journal of Web Semantics 2018; 49: 1-15.

[57] Masthoff J. Group Recommender Systems: Combining Individual Models. In: Ricci F, Rokach L, Shapira B, Kantor P (editors). Recommender Systems Handbook. Boston, MA, USA: Springer, 2011. pp. 677-702.

[58] Arroyo S, Ding Y, Lara R, Stollberg M, Fensel D. Semantic web languages. strengths and weakness. In: Applied Computing (IADIS04) Conference; Lisbon, Portugal; 2004. pp. 23-26.

[59] Suchanek FM, Kasneci G, Weikum G. Yago: A core of semantic knowledge. In: World Wide Web Conference; Alberta, Canada; 2007. pp. 697-706.