An Open Framework for Multi-source, Cross-domain Personalisation with Semantic Interest Graphs[*]

Benjamin Heitmann
Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway, Ireland
benjamin.heitmann@deri.org

ABSTRACT

Cross-domain recommendations are currently available in closed, proprietary social networking ecosystems such as Facebook, Twitter and Google+. I propose an open framework as an alternative, which enables cross-domain recommendations with domain-agnostic user profiles modelled as semantic interest graphs. This novel framework covers all parts of a recommender system. It includes an architecture for privacy-enabled profile exchange, a distributed and domain-agnostic user model and a cross-domain recommendation algorithm. This enables users to receive recommendations for a target domain (e.g. food) based on any kind of previous interests.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering

General Terms

Architecture, Algorithm, User model

Keywords

Personalisation, Cross-domain, Multi-source, Domain-agnostic

1. INTRODUCTION

Personalised recommendations have proven themselves to greatly enhance the user experience of searching, exploring and prioritising new and interesting content on social networking sites, like Facebook, Twitter and Google+. However all of these social networking sites are also each the centre of an ecosystem which allows external services and 3rd party sites to enhance the user experience. Music streaming services, news papers or image sharing sites can post to a users activity stream on behalf of the user, thus contributing interests to his profile.

While this increases the value of the social network for the user and results in stronger user attachment, this also introduces completely new challenges for building recommender systems. The input data consists of user profiles which are (a) distributed across different services (*multi-source profiles*), and (b) not associated with a specific domain or inventory (*domain-agnostic profiles*) [1]. A source of background knowledge which provides connections between interests is required. Finally, the recommendation algorithm needs to provide *cross-domain recommendations* [8] targeted to a specific domain and inventory, in order to e.g. recommend posts from a food blog to users without food interests, or select advertisements from a pool of ads.

In order to provide users with the benefit of multi-source, cross-domain recommendations, users currently need to accept a trade-off [3] regarding their *privacy, trust, data ownership and control*: Privacy and personalisation are currently at odds [11]. Enabling 3rd party services requires sharing of user data with external services, thus introducing the potential for data leaks. In addition, the social networking operator, e.g. Facebook, owns all of the user data. The user has no control to whom his data is sold.

The goal of my research is to provide the same recommendation capabilities, namely multi-source and cross-domain recommendations, outside of closed social networking ecosystems. Towards this goal I am developing an open framework, as illustrated in figure 1. It addresses fundamental issues in the architecture and methodology of making recommendations on heterogeneous, distributed data. It consists of (i) an architecture for exchanging and aggregating of user profile data, (ii) user model, (iii) background knowledge and (iv) recommendation algorithm. In my PhD thesis, I plan to make contributions to the following three research areas:

Architecture for privacy-enabled profile exchange: Can user profile data and user profile fragments be exchanged in a decentralised way, while protecting the privacy of the users?

Distributed and domain-agnostic user model: What kind of data structure is required in order to allow merging profile fragments from multiple sites? What background knowledge can be used to build domain-agnostic user profiles? How can domains and genres and their constituent entities be defined in a flexible and universal way?

[*]The work presented in this paper has been funded in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lìon-2) and by the ADVANSSE project funded by Cisco Systems. In addition, I would like to thank my PhD supervisor Dr. Conor Hayes as well as Dr. Maciej Dabrowksi for their support and feedback.

RecSys'12, September 9–13, 2012, Dublin, Ireland.

Cross-domain recommendation algorithm: Which class of recommendation algorithm can process domain-agnostic user profiles in order to provide recommendations for a specific target domain and inventory? Which data sets and metrics can be used for evaluating the performance of cross-domain recommendations?

2. BACKGROUND

I propose a framework, which addresses shortcomings in the way that current social networking sites provide cross-domain recommendations. Namely, I address the closed nature of current social network ecosystems, the missing portability of user profiles and interests, and the proprietary and secret recommendation algorithm.

In social networking ecosystems one site typically has the role of the *hub site*, which provides the main entry point for the whole ecosystem and stores the user profile data. Prominent hub sites are the social networking sites Facebook, Twitter and Google+. *Third party services* can provide value-added and personalised services for the user of an ecosystem.

In order to give the user a sense of privacy, while aggregating his profile data from external services, current social networking ecosystems are fundamentally *closed* [6]: The user model and all aggregated data about a user belong to the social network operator, e.g. Facebook. This is meant to protect the privacy of the user when his profile data is exchanged between the hub site and an external 3rd party service. However, it also creates user lock-in and data silos [3]. Users can not control to whom their data is sold, as the data generated by the users is the actual product. As an alternative, I propose an open architecture for exchange of user profile fragments in section 3. It uses open standards to enable a decentralised ecosystem in which users are not dependent on social network operators for recommendations.

All major social networking sites cater to general purpose sharing, as they are not limited to a particular genre or domain (like music or sports) or to a specific inventory (such as books in stock at a physical warehouse). This results in domain-agnostic user profiles. However, these user profiles are *not portable* between social networks [1], as they use identifiers for the interests which are specific to each social network operator. In addition, it is difficult to merge user profiles from different sources. Lastly, different social networks have different ways of classifying interests, so that it is difficult to specify a "food" recommendation in a universal way. In section 4, I present a distributed and domain-agnostic user model based on semantic interest graphs. It enables portability and merging of user profiles, as well as the definition of domains to categorise interests.

Facebook provides a so-called "social plugin" for cross-domain recommendations, which allows an external site to recommend content or items from its inventory to users without prior interest in the domain of the external site. E.g. a weblog about rugby, can use this plugin to recommend its articles to users without prior preferences about "sports". However, Facebook's cross-domain recommendations use *a proprietary algorithm*, so that users without a Facebook account can not benefit from this new kind of personalisation. In section 5, I introduce a graph-based algorithm which can use semantic interest graphs and domain definitions from section 4 to provide recommendations from a target domain. In addition, I introduce the challenges for evaluating the user utility and performance of cross-domain recommendation in section 6.

3. ARCHITECTURE FOR PRIVACY-ENABLED PROFILE EXCHANGE

We propose a novel, open architecture as an alternative to closed architectures for exchanging and aggregating user profile data. An architecture prescribes (a) the standards as well as (b) the roles and (c) the communication pattern between the different participants. This makes it possible to align the different interests of all stakeholders. By implementing an architecture, all *individual* participants agree on the same technical principles, which in turn allows the architecture to guarantee the identified requirements on a *global* level.

Requirements: The architecture must support a federated and scalable ecosystem with any number of hub sites, 3rd party services and users. The architecture must enable interoperability of user profile data between 3rd party services and hub sites. At the same time, user data must not be allowed to leak to any unauthorised parties. These could be other web sites, adversaries or just other users. In other words, the architecture must be privacy-enabling and interoperable *at the same time.*

Our architecture prescribes the usage of the following standards: Linked Data [2], and the Friend-of-a-Friend (FOAF) and Semantically Interlinked Online Communities (SIOC) vocabularies allow the description of domain-agnostic user profiles. WebIDs securely connect a user identity to the information in a user profile and can be used for authenticating a user. The WAC vocabulary allows the user to authorise third party services for accessing different parts of his profile information.

The interplay between FOAF, WebIDs and the WAC vocabulary requires the participants to perform one of three roles: profile storage services, data consumers and user agents. We specify a communication pattern, which prescribes how the participants need to interact with each other according to their role. This communication pattern is described in full in [6], as well as a qualitative evaluation of the presented architecture based on the evaluation framework for privacy-enhanced personalisation suggested by Wang and Kobsa [11].

We summarise the qualitative evaluation as follows: The architecture provides a universal ecosystem, as all participants will support the same standards and implement the same communication pattern. The architecture is scalable, as there are no bottlenecks or central points of failure, due to the decentralised nature of the used standards. WebIDs allow anybody to host any number of identities on any server, and the WAC vocabulary allows authorising the access of a resource on the same server on which it is stored. For profile storage and data consumption existing standards and infrastructure from the World Wide Web and the Web of Data, such as HTTP and RDF are reused, thus making future adoption by service providers easy.

4. DISTRIBUTED AND DOMAIN-AGNOSTIC USER MODEL

The presented architecture enables aggregating domain-agnostic user profiles from multiple sources. In order to use

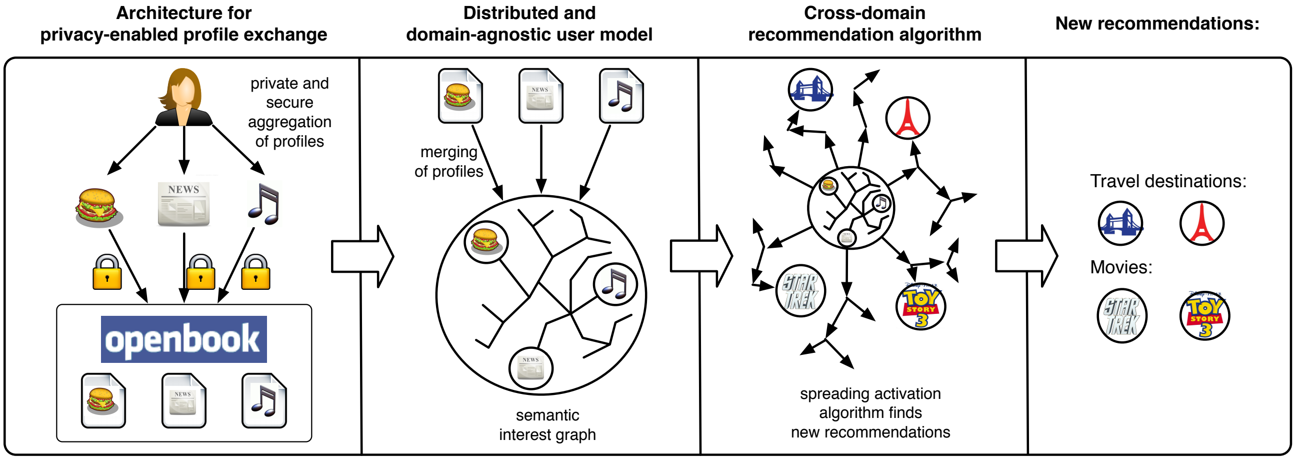
Architecture for privacy-enabled profile exchange
Distributed and domain-agnostic user model
Cross-domain recommendation algorithm
New recommendations:
private and secure aggregation of profiles
NEWS
openbook
merging of profiles
semantic interest graph
spreading activation algorithm finds new recommendations
Travel destinations:
Movies:
STAR TREK
TOY STORY 3

Figure 1: Overview of the proposed open framework for multi-source, cross-domain personalisation

these profiles for personalisation, a novel user model which fulfils the following requirements is needed.

Requirements: The data structure of the user model needs to allow merging of aggregated profiles from different sources. It must be able to model interests from many diverse domains, as well as the connections between interests. The representation of the interests must not be dependent on a specific inventory of items, i.e. it must not use identifiers which only exist at one data source. In order to utilise the user model for recommendations, a source of background knowledge which provides connections between interests is required. Finally, in order to provide cross-domain recommendations, it must be possible to define the target domain and its constituent items. These domain definitions are required in order to differentiate domains, or to use preferences in one domain for recommendations in another domain.

We propose using semantic interest graphs to represent the user profiles. Traditional user profile representations are not suitable to the requirements. *Vectors of item ratings* can not be exchanged between sites with different inventories, and they can not represent interests outside of the inventory. *Lists of plain text items*, which can represent tags or frequent words, can represent any kind of interest independent of the inventory. But they provide only limited support for finding relationships between interests.

In a *semantic interest graph*, each interest is represented by an RDF entity. RDF uses a graph data model [2], which allows RDF entities to be linked to each other. By leveraging existing knowledge engineering from the Linking Open Data community project, vocabularies and ontologies which already define any kind of interests can be reused. We propose to use concepts from the DBpedia knowledge base as *background data*, which is an RDF version of Wikipedia, and provides RDF concepts for any page or category from Wikipedia. In addition, each DBpedia concept is linked to other DBpedia concepts, which allows discovering relationships between concepts.

Using DBpedia concepts to represent user profiles as semantic graphs, enables exchanging and merging of profile data and independence from specific domains and inventories. The concept identifiers stay the same, independent of the site on which the profile was created. Profiles can be easily merged when all sites use identifiers from DBpedia. Any concept from Wikipedia can be used as interest, independent of any domain or inventory restrictions.

In order to define domains and genres in a flexible and universal way, we propose to use the reasoning capabilities of RDF Schema and SKOS. RDF Schema allows defining hierarchies of classes and class instances. The Simple Knowledge Organisation System (SKOS) vocabulary allows defining relationships between categories and topics, such as broader or narrower categories. Together, RDF Schema and SKOS allow defining of reasoning rules to determine which entities belong to a domain. E.g. Pizza is categorised as WorldCuisine which has a super-category of Food, so it belongs to the Food domain.

Open issues: What process is required for specifying domains and entity types, and can this be automated based on e.g. the inventory of a site? How should the rule based reasoning process and the recommendation process be interleaved in order to support cross-domain personalisation?

5. CROSS-DOMAIN RECOMMENDATION ALGORITHM

After exchanging user profile data via the proposed architecture, and merging them to an interest graph with the proposed methodology, an appropriate recommendation algorithm is required. Cross-domain recommendation is a novel personalisation approach, which can utilise user preferences in one domain (e.g. music) to suggest recommendations in another domain (e.g. movies) [8].

Requirements: The algorithm must be graph-based, as both the input data and the background data for the algorithm are both graph-based. The algorithm must be able to take the semantics of the graph into account, e.g. by differentiating between different types of connections or different entity types. In addition, it must be able to differentiate between items from a target domain and from other domains.

We propose using a spreading activation (SA) algorithm as a recommendation algorithm. Spreading activation is inspired by the fact that human memory retrieves memories by association. Crestani [4] describes the algorithm as follows: An activation value spreads to all direct neighbours of the starting nodes. As soon as the activation threshold of a node is reached, it counts as activated. Activated nodes are used as the starting points for the next activation phase. *Unconstrained* SA will quickly cover the whole graph, however the activation can be constrained. Possible *constraints* include the distance to the start nodes, the number of outlinks of a node (fan-out constraint), the type of the link (path constraint), or the type of activated nodes (activation constraint).

In order to use spreading activation to provide recommendations in a target domain, the path and activation constraints can be used. Based on the target domain definition, the weights of all link types which belong to entities from the domain are set higher then other link types. In addition, the spreading step is repeated until enough nodes from the target domain are activated, using the activation constraint. This allows e.g. recommending book authors to a user who has only travel interests and destinations in his user profile.

Open issues: The algorithm can be parameterised in different ways, however the effect of e.g. the fan-out penalty on the results need to be measured using the performance metrics of the overall evaluation framework. An additional possibility is performing a user based evaluation.

6. EVALUATION

We are planning to evaluate the presented framework for multi-source, cross-domain personalisation as part of the ADVANSSE [5] collaboration project with Cisco Ireland. ADVANSSE has the goal of providing personalisation embedded in the IT landscape of current enterprises, which are characterised by large and heterogeneous information systems and social platforms. This results in user profiles being spread across multiple systems and user interests from many diverse domains.

Requirements for the ADVANSSE prototype: The implementation must be very scalable, in order to provide sufficient performance on large datasets from enterprise IT systems. In order to achieve this, the implementation should distribute the processing among multiple servers.

In order to provide a very scalable implementation, we are implementing our spreading activation algorithm with the Apache Giraph Java library for large-scale graph processing. Giraph runs on top of Apache Hadoop, which is an implementation of the MapReduce framework. As background data we use a subset of DBpedia which consists of 11 million entities and 40 million edges. For evaluation purposes we use user profiles from the question answering sites of the StackExchange network. The StackExchange network consists of sites from very different domains, such as StackOverflow for computing questions, cooking, photography and bicycle sites.

Viewing the algorithm as a link predication task provides a general evaluation framework [9], in which area under the curve (AUC) and precision can be used as performance metrics. Comparable baseline approaches are provided by Linked Data Semantic Distance (LDSD) [10] and Random Walk with Restart (RWR) [7].

Open issues: The execution speed of the algorithm can be improved by pre-selecting a smaller subset of the data. However what is the trade-off in user utility of the personalisation results?

7. CONCLUSIONS AND FUTURE WORK

In this extended abstract I have presented an overview of my PhD research. My goal is to enable cross-domain recommendations from multi-source user profiles outside of closed social networking ecosystems. Towards this goal I am developing an open framework, which covers all parts of a recommender system. I propose an *open architecture* for profile data exchange as an alternative to current, closed architectures. It empowers users by using open standards to protect their privacy. I present a *distributed and domain-agnostic user model* to represent the user interests based on Linked Open Data standards and DBpedia concepts. It enables merging of profile fragments and the definition of domains to categorise interests. *Spreading activation* is a graph-based and content-based recommendation algorithm, which can provide cross-domain recommendations.

The goal of my final year research will be a systematic evaluation of the performance and user utility of cross-domain recommendations, as both the execution speed and the user utility depend on the parameterisation of the algorithm.

8. REFERENCES

[1] F. Abel, E. Herder, G. Houben, N. Henze, and D. Krause. Cross-system user modeling and personalization on the social web. *User Modeling and User-Adapted Interaction (UMUAI)*, 22(3):1–42, 2011.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[3] R. Chellappa and R. Sin. Personalization versus Privacy: An Empirical Examination of the Online Consumers Dilemma. *Information Technology and Management*, 6(2):181–202, 2005.

[4] F. Crestani. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6):453–482, 1997.

[5] B. Heitmann, M. Dabrowski, A. Passant, C. Hayes, and K. Griffin. Personalisation of Social Web Services in the Enterprise Using Spreading Activation for Multi-Source, Cross-Domain Recommendations. In *AAAI Spring Symposium on Intelligent Web Services Meet Social Computing*, 2012.

[6] B. Heitmann, J. G. Kim, A. Passant, C. Hayes, and H.-G. Kim. An architecture for privacy-enabled user profile portability on the Web of Data. In *Int. Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2010)*, 2010.

[7] G. Jeh and J. Widom. Scaling personalized web search. In *World Wide Web Conference*, 2003.

[8] A. Loizou. *How to recommend music to film buffs: enabling the provision of recommendations from multiple domains*. PhD thesis, University of Southampton, 2009.

[9] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.

[10] A. Passant. dbrec – music recommendations using dbpedia. *ISWC*, 2010.

[11] Y. Wang and A. Kobsa. Technical Solutions for Privacy-Enhanced Personalization. *Intelligent User Interfaces: Adaptation and Personalization Systems and Technologies*, 2009.