Leveraging Linked Data Analysis for Semantic Recommender Systems

Andreas Thalhammer

STI Innsbruck, University of Innsbruck, Austria
andreas.thalhammer@sti2.at

1 Motivation

Traditional (Web) link analysis focuses on statistical analysis of links in order to identify "influencial" or "authorative" Web pages like it is done in PageRank, HITS and their variants [10]. Although these techniques are still considered as the backbone of many search engines, the analysis of usage data has gained high importance during recent years [12]. With the arrival of linked data (LD), in particular Linked Open Data (LOD),[1] new information relating to *what* actually connects different vertices is available. This information can be leveraged in order to develop new techniques that efficiently combine linked data analysis with personalization for identifying not only relevant, but also diverse and even missing information. Accordingly, we can distinguish three problems that motivate the topic of this thesis:

Relevance. LOD is well known for providing a vast amount of detailed and structured information. We believe that the information richness of LOD combined with user preferences or usage data can help to understand items and users in a more detailed way. In particular, LOD data can be the basis for an accurate profile which can be useful for recommendation in various domains. As information about items in the user profile is often unstructured and contains only little background knowledge, this information needs to be linked to external sources for structured data such as DBpedia.[2] Also, product and service providers need to link their offers accordingly. Methods for this two-way alignment have to be specified and evaluated.

Diversity. According to [5], recent developments in semantic search focus on contextualization and personalization. However, approaches that semantically enable diverse recommendations for users, also in context to users' profiles, remain barely explored. Of course, this states a complementary way of recommendation that is often only based on ranking by relevance. Consider the example of a news aggregation Web site which ranks articles by popularity. Popular articles are placed at the main page. On the same topic, there are hundreds of additional articles from other news sites and blogs indexed, but not visible on the main page. Of course, these articles get much

[1] Linking Open Data - http://ow.ly/8mPMW
[2] DBpedia - http://dbpedia.org/

E. Simperl et al. (Eds.): ESWC 2012, LNCS 7295, pp. 823–827, 2012.

less clicks than the ones from the main page. This results in the most popular sites gaining even more popularity. Our goal is to break up such feedback loops by introducing diverse recommendations.

Non-existence. During recent years the LOD cloud[3] has been growing to a huge amount of interconnected triples. Approaches like Freebase[4] and the Wikidata project[5] focus on collecting information through direct input in order to establish an encyclopedic corpus. We believe that these systems are likely to face the same issue like Wikipedia:[6] since 2006, its growth is decreasing [16]. This problem of Wikipedia gained focus of research and different article recommendation systems have been explored [4,8]. These systems point users to articles they might want to edit. This idea can be extended in order to work for Freebase or Wikidata: given the corresponding user profiles, it becomes feasible to point users to missing facts (e.g. the mayor of a city).

2 Related Work

During the last decade, several approaches that aim to link semantic technologies and recommender systems have been introduced. [15] introduces a framework that enables semantically-enhanced recommendations in the cultural heritage domain. Recommendation as well as personalization in this work rely on the CHIP ontology which is designed specifically for the cultural heritage domain. The core of the recommendation strategy bases on discovering domain-specific links between artworks and topics (e.g. the same creator, creation site, or material medium). In the outlook section of this work, the author emphasizes on LD as a core technology to enhance personalization and recommendation. The work presented in [7] states an approach to utilize LD in order to enhance recommender systems. Just like our focus on linking items in the user profile to LOD items, Heitmann and Hayes utilize LOD links in order to enhance background information for recommendation corpora. The recommendation approach is collaborative filtering (cf. [1]) as "the inconsistent use of these semantic features makes the cost of exploiting them high" [7]. In my thesis, I try to leverage exactly these semantic features for recommendation. [2] introduces a semantic news recommender system called "News@hand" that makes use of ontology-based knowledge representation in order to mitigate the problem of ambiguity and to leverage reasoning for mediating between fine and coarse-grained feature representations. The system supports content-based as well as collaborative recommendation models. Similar to our approach, the items and the user profiles are represented as a set of weighted features. The weights of the item features are computed with a TF-IDF technique which does not involve additional knowledge.

[3] LOD cloud - http://lod-cloud.net/
[4] Freebase - http://www.freebase.com/
[5] Wikidata - http://meta.wikimedia.org/wiki/Wikidata
[6] Wikipedia - http://wikipedia.org/

DBrec, a music-specific recommendation approach, is presented in [13]. In this paper, LD (i.e. DBpedia) is utilized for semantic recommendations. The approach is based on the similarity measure LDSD which aims at estimating the distance between two LD resources by considering entities in a two-hop radius. Just like in item-based collaborative filtering (cf. [1]), the similarity scores between entities can be computed offline. For storing the results, the author introduces the LDSD ontology. In comparison to the similarity measures we try to introduce, LDSD similarity computation is based on statistics about direct and indirect in and out links disregarding the importance of particular predicates.

Personalized property suggestions relate to two topics, i.e. personalized article recommender systems and property suggestion systems. Several recommender systems utilize user information to make recommendations like "You might want to edit this Wikipedia article" [4,8]. Others suggest facts that might be relevant for a certain Wikipedia entry [11,17]. However, the problem of proposing properties for Wikipedia, Freebase, or OntoWiki documents in a personalized way, like "Here are some missing facts about [some article] that you could know", has not been addressed so far.

3 Proposed Approach

Our approach is based on identifying distinctive item features with the help of usage or rating data. As with all recommender systems, the main goal is to help users to find information that is important to them. On a different level, the macro goals are to identify and match information that is important about users with information that is important about items. Accordingly, the first part of the process can be broken down to the following steps:

1. Extract and create user profiles with items linked to LOD.
2. Find k-nearest neighbors for each item according to the user-item matrix. The entries of this matrix are ratings by users for items (cf. [14]).
3. Identify which features are important about specific items (which property-value pairs do they share with their neighbors).
4. Combining the results step 1 and 3, it is possible to find out what is specific about a user.

These processing steps can be performed offline. For the representation and storage of the results an appropriate vocabulary needs to be selected. After these steps, we have established a situation where we know what is important about specific items as well as users. Afterwards, in the second part, we investigate for different match-making techniques that help us to recommend items to the user that relate to relevant, diverse, or missing information. Accordingly, the following techniques may serve as starting points:

Relevance. Graph matching of the weighted user and item graphs.
Diversity. Clustering according to the most important properties or property-instance combinations in the user profile.

Non-existence. Recommend a missing fact which is very important for an article (e.g. population of a city) to the user who is frequently editing this property in related articles.

Some efforts towards semantically enabled cross-domain profiles and recommendations that refer to LOD have already started [6]. However, utilizing user profile or usage data in order to introduce feature weights for items has not been explored to the best of my knowledge. Also, using collaborative filtering background knowledge in order to discover important properties of LOD entities is a novel approach. There might also be a limitation of the approach: The user-item matrix that we make use of can be used directly for collaborative filtering. In previous experiments that were conducted by various users in connection with the Netflix Prize[7] it turned out that sole collaborative filtering outperformed all approaches that tried to enrich the data set with background information. The pure collaborative filtering approach was much better in performance and also in terms of prediction quality. However, we target scenarios that are not constrained by a single fixed domain (such as movies). Moreover, measuring relevance only is not comparable to our scenarios as we also try to incorporate diverstiy. Given a specific domain, collaborative filtering can serve as a base line for our relevance-focused approach.

4 Methodology

For my thesis, I will conduct the following steps:

- identify related fields of research
- implement linking of items to LOD and the according weighted item and user representations
- identify existing match-making algorithms and evaluate their appropriateness
- conduct different algorithms for match-making
- evaluate the selected approaches according to relevance, diversity and estimated information gain

For first tests and results, we chose the HetRec2011 MovieLens2k dataset [3] that has been linked to Freebase data. The rating data stems from the MovieLens10M dataset[8] that contains anonymous user profiles.

The evaluation of recommender systems is usually based on precision and recall which can also be applied in this case. In this field, a couple of approaches already exist that can serve as base lines. A measure which ranks diversity is introduced in [9]. The recommendation of missing content for Web 2.0 collections can be evaluated by comparing the number of edits with and without the recommendation approach.

[7] Netflix Prize - http://www.netflixprize.com/
[8] MovieLens10M - http://www.grouplens.org

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17, 734–749 (2005)
2. Cantador, I., Bellogín, A., Castells, P.: Ontology-based personalised and context-aware recommendations of news items. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2008, vol. 1, pp. 562–565. IEEE Computer Society, Washington, DC (2008)
3. Cantador, I., Brusilovsky, P., Kuflik, T.: 2nd ws. on information heterogeneity and fusion in recommender systems (hetrec 2011). In: Proc. of the 5th ACM Conf. on Recommender Systems, RecSys 2011. ACM, New York (2011)
4. Cosley, D., Frankowski, D., Terveen, L., Riedl, J.: SuggestBot: Using Intelligent Task Routing to Help People Find Work in Wikipedia. In: Human-Computer Interaction (2007)
5. Dengel, A.: Semantische suche. In: Dengel, A. (ed.) Semantische Technologien, pp. 231–256. Spektrum Akademischer Verlag (2012)
6. Fernández-Tobías, I., Cantador, I., Kaminskas, M., Ricci, F.: A generic semantic-based framework for cross-domain recommendation. In: Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems, HetRec 2011 (2011)
7. Heitmann, B., Hayes, C.: Using Linked Data to Build Open, Collaborative Recommender Systems. Artificial Intelligence (2010)
8. Huang, E., Kim, H.J.: Task Recommendation on Wikipedia. Data Processing (2010)
9. Murakami, T., Mori, K., Orihara, R.: Metrics for Evaluating the Serendipity of Recommendation Lists. In: Satoh, K., Inokuchi, A., Nagao, K., Kawamura, T. (eds.) JSAI 2007. LNCS (LNAI), vol. 4914, pp. 40–46. Springer, Heidelberg (2008)
10. Ng, A.Y., Zheng, A.X., Jordan, M.I.: Stable Algorithms for Link Analysis. Machine Learning, 267–275 (2001)
11. Oren, E., Gerke, S., Decker, S.: Simple Algorithms for Predicate Suggestions Using Similarity and Co-occurrence. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 160–174. Springer, Heidelberg (2007)
12. Pariser, E.: The filter bubble: what the Internet is hiding from you. Viking, London (2011)
13. Passant, A.: dbrec — Music Recommendations Using DBpedia. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part II. LNCS, vol. 6497, pp. 209–224. Springer, Heidelberg (2010)
14. Sarwar, B., Karypis, G., Konstan, J., Reidl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web, WWW 2001, pp. 285–295. ACM, New York (2001)
15. Wang, Y.: Semantically-Enhanced Recommendations in Cultural Heritage. PhD thesis, Technische Universiteit Eindhoven (2011)
16. Wikipedia. Modelling wikipedia's growth, `http://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth` (online accessed March 12, 2012)
17. Zhang, H., Fu, L., Wang, H., Zhu, H., Wang, Y., Yu, Y.: Eachwiki: Suggest to be an easy-to-edit wiki interface for everyone. In: Semantic Web Challenge (2007)