AAAI Technical Report SS-12-04
Intelligent Web Services Meet Social Computing

Personalisation of Social Web Services in the Enterprise Using Spreading Activation for Multi-Source, Cross-Domain Recommendations*

Benjamin Heitmann and **Maciej Dabrowski**
Alexandre Passant and **Conor Hayes**
Digital Enterprise Research Institute
National University of Ireland, Galway
Galway, Ireland
firstname.lastname@deri.org

Keith Griffin
Cisco Systems
Galway, Ireland
kegriffi@cisco.com

Abstract

Existing personalisation approaches, such as collaborative filtering or content based recommendations, are highly dependent on the domain and/or the source of the data. Therefore, there is a need for more accurate means to capture and model the interests of the user across domains, and to interlink them in a semantically-enhanced interest graph. We propose a new approach for multi-source, cross-genre recommendations that can exploit the heterogeneous nature of user profile data, which has been aggregated from multiple personalised web services, such as blogs, wikis and microblogs. Our approach is based on the Spreading Activation model that exploits intrinsic links between entities across a number of data sources. The proposed method is highly customizable and applicable both to generic and specific recommendation scenarios and use cases. With the growing number of Social Web applications in the enterprise (blogs, wikis, micro blogging, etc.), it becomes difficult for knowledge workers to avoid content overload and to quickly identify relevant people, communities and information. We demonstrate the application of our approach in an industrial use case that involves recommendation of social semantic data across multiple services in a distributed collaborative environment.

1 Introduction

In the context of modern enterprises the employees are often distributed across departments and geographical locations, use different information systems, and share skills and expertise that spans multiple knowledge domains. While users of the Social Web have come to expect personalised services for various types of content such as music, books or social activity streams, personalisation for enterprise users remains a challenge due to factors such as distribution of users and the fact that their expertise can span multiple domains. The provision of social platforms that enable personalized information access in an enterprise environment involves an infrastructure and algorithms that address many complex challenges. First, aggregation and integration of data from different information systems/social platforms in the IT landscape of the enterprise is required. Second, existing personalisation methods, such as collaborative filtering or content based recommendations, are highly dependent on the domain and/or the source of the data. Therefore, there is a need for an approach that exploits cross-domain data from different sources. Such a method requires more accurate means to capture and model the interests of the user across different domains, in order to interlink the interests of the user and create the full interest graph of the user out of the profile fragments across distributed social platforms. This interest graph can then in turn be used for recommending more accurate and context-dependant resources available in the enterprise social platform, mitigating the threat of information overload and enabling discovery of valuable information and/or people. This paper provides an overview of the state-of-the-art in personalization methods and introduces a new approach for personalisation in an enterprise context based on Spreading Activation (SA).

Our approach is based on the Spreading Activation model that exploits available metadata (e.g. semantic tags) to provide rich recommendations using the intrinsic links between entities across data sources. The method enables personalization beyond the context of a single web service (i.e. social platform). We argue that the method scales well as it utilizes the local properties of the graph, that is direct neighbourhood links, in contrast to approaches such as PageRank that use global properties. Further, no pre-processing is required to compute recommendations assuming availability of the interest graph. The proposed method is highly customizable and applicable both to generic and specific recommendation scenarios and use cases. While it is applicable to multi-domain data to produce generic recommendations, it can also be customized towards a particular domain to enable domain-specific recommendations.

In this paper we give an overview of the personalisation approaches applicable to enterprise social networks in general, with a particular focus on graph-based personalisation approaches. Based on the enterprise usecase (see Section 1.1), we derive the requirements for personalisation (see Section 2) and give an overview of the most popular personalisation approaches (see Section 3), including a detailed description of the spreading activation algorithm. Finally, we present how the discussed methods can be applied in the enterprise usecase (see Section 4) and conclude the paper

*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.
Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

with the summary of contributions and future work (see Section 5).

1.1 Usecase

Andrew, Bob, and Cecilia are experts in Semantic Web technologies employed by the same organization (see Figure 1) working in different departments (or locations). Andrew (CTO) is interested in the Semantic Web. Therefore, Andrew has to follow corporate blogs of Bob and Cecilia (both interested in Semantic Web) to discover any content on that topic published in the organization. Currently, this task requires Andrew to log-in to three different social spaces to access relevant content. In addition, as Bob is a new user, Andrew may not be aware of Bob's presence in the system and thus won't not able to discover content published by Bob. Further, Andrew would like to receive instant updates in his social dashboard as soon as new content is published. Although approaches such as RSS seem suitable, they imply regular querying of the information sources for updates across corporate networks with restricted access policies.

Another use case involves the use of different terminology to describe same/similar topic across departments. While Bob and Cecilia may use different terms to express their expertise and interests, there is a strong implicit relation between the concepts in their respective user profiles (Semantic Web, Databases). Nevertheless, a report about a new product published by Bob may be not visible to Cecilia due to a number of reasons such as the use of separate social platforms or no direct match between the resource annotation (tags) and user interests. This second use case drives the majority of the requirements for personalisation of social content in a distributed enterprise context, which are explained in more detail in the next section.

2 Requirements for personalization

The section discusses the requirements for a personalization approach applicable to social platforms in large and distributed organization that were derived from the two use cases (see Section 1.1).

Mitigation of the cold-start problem: The cold start problem refers to the fact that the quality of recommendations gets worse, if there is no data available about either the user or the recommendable items. The personalisation approach should be able to mitigate this problem by using user/item properties available either in the system or in other (external) sources.

Avoid lack of recommendations: The lack of recommendations is a special case of the cold-start problem as the recommendation algorithm should provide a continuous stream of relevant and ranked recommendations, so that there is no lack of recommendations in the user interface. The problem involves a trade-off between the real-time performance of the algorithm and the pre-computation of potential recommendations, as well as the number of available recommendations and their quality/accuracy.

Scalability: The approach should be scalable and applicable to large amounts of data. Ideally, the performance of the method should not depend on global properties, such as the total number of users and/or recommendable item, but on the density of the integrated graph in the neighbourhood of a user or an item.

Multi-source recommendations: The method must be applicable to data from multiple sources and described with various standards (e.g. spread across multiple RDF stores using different metadata standards).

Cross-domain recommendations: The method should be applicable to data from different genres/domains. Topical variety of documents in the enterprise can be vast: from technical reports that discuss low level technical aspects to marketing plans or budget sheets. The method should exploit data available about users/items across domains to make higher quality recommendations.

Universality: The recommendation approach should be independent of the source data format. The source data might be available in different formats, such as emails, instant messages or office documents and spreadsheets.

Recommendation of sources and items: The use cases suggest the need for two different recommendation scenarios: recommendation of items that fit the interest profile of a user and discovery of sources of items of potential interest, such as users or discussion groups.

Customization for a specific domain: The personalization approach should allow the provision of generic (topic-independent) recommendations as well as customization for topic-specific suggestions. For example, for a given recommendation context it may be important if people collaborated in the same projects, but not that they play in the same office sports team.

3 Personalization approaches

Personalised recommendations greatly enhance the user experience of searching, exploring and finding new and interesting content (Montaner, López, and De La Rosa 2003), however, mostly using homogeneous data from one source. In contrast, the use of personalisation in a distributed enterprise context is more challenging, as the personalisation has to take multiple sources, formats and genres into account.In this section we first introduce a classification of the most well established personalisation approaches.

3.1 Classification of personalisation approaches

The typical recommender systems require three components to provide recommendations (Burke 2002): (1) background data, which is the information the system has before the recommendation process begins, (2) input data, also called the user model, which is the information provided about the user in order to make a recommendation, and (3) the recommendation algorithm which operates on background and input data in order to provide recommendations for a user. Based on these attributes the recommendation approaches can be grouped in 4 classes (Burke 2002): Collaborative filtering, content-based recommendation, knowledge-based recommendation and hybrid approaches. We also describe

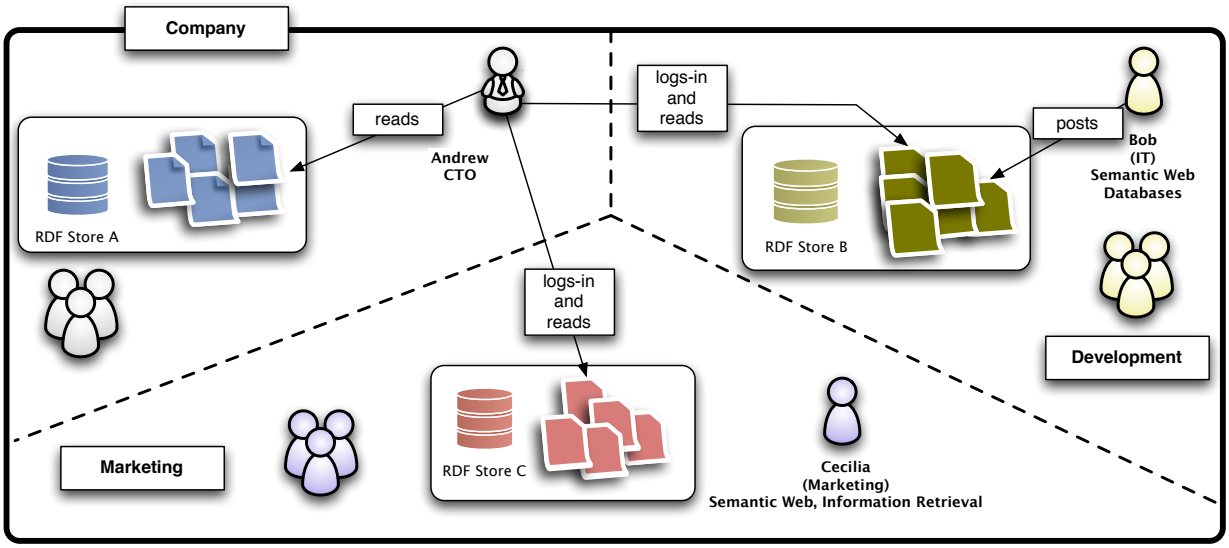

Figure 1: Current state of the Social Semantic Enterprise systems divided into many division-specific platforms.

what is commonly called the "top-k approach", which however refers to a term from the database community. In addition, we introduce two graph-based recommendation approaches.

Collaborative filtering Collaborative filtering aggregates ratings for items from different users, and uses similarities between users to recommend items. It is probably the most mature and widely implemented recommendation algorithm, because it achieves fairly good results and is easy to implement (Burke 2002). It only requires data about the ratings between users and items as background data, no other information about either the users or the items is required (Herlocker et al. 2004). The input data usually consists of a user profile providing ratings for one or more items. The recommendation algorithm uses the background data to calculate the pair-wise similarity between all items or all users, and then uses the input data to recommend similar users or items (Sarwar et al. 2001).

Content-based recommendation Content-based recommendation approaches use features of the items as the background data for the recommendation (Pazzani and Billsus 2007). These can either be directly derived from the content, e.g. keywords from text or tempo of the music piece, or derived from the meta-data of the items, e.g. author, title and genre. The input data needs to describe the users preferences in terms of content features. Both the background and the input data require the consistent description of content features in order to match the user preferences to the features of the content.

Knowledge-based recommendation Knowledge-based recommendation approaches aims to suggest items based on inferences about the users' needs and preferences. This requires background data that includes knowledge about users and items, which is sufficient in consistency and scale for making inferences. The input data needs to provide knowledge about the users needs and preferences which can be mapped to the knowledge about users and items in the background data. Knowledge-based approaches are distinguished in that they have functional knowledge, e.g. about how a particular item meets a particular user need, and can therefore reason about the relationship between a need and a possible recommendation (Burke 2002).

Amini et al. (Amini, Ibrahim, and Othman 2011) provide an overview of current approaches for knowledge-based recommendation, and suggest that the most important way to apply knowledge to the personalisation task lays in expressing the context of the user.

Hybrid algorithms Hybrid algorithms combine two or more recommendation algorithms to provide better results with fewer of the drawbacks of an individual algorithm. In order to combine two algorithms different methods can be used (Burke 2002), e.g.: the scores of several algorithms can be combined with weights; the output of one algorithm can be used as the input for the next one, thus forming a cascade; the system can switch between different algorithms depending on the situation; the presentation of the output of several algorithms can be mixed in the user interface. Most commonly, collaborative filtering is combined with an algorithm of a different type, e.g. a content-based one, in order to mitigate situations in which not enough background data for an item or a user is available.

Top-K approach The term "Top-k" refers to a database query which retrieves the first (i.e. best) k results for a query. Usually the ranking of the query is based on a scoring function, which utilizes different properties of the entities that are queried (Ilyas, Beskales, and Soliman 2008) The view present in the recommender systems research community

views "Top-k" as a problem of identifying a set of N items of highest interest to a user, also called *top-N recommendation problem* (Deshpande and Karypis 2004). All previously discussed classical approaches for building recommender systems can be used for the top-N recommendation problem by ranking the recommendation results and returning only the first N results. However, in contrast to top-k queries, top-N recommendations have an explicit user model which heavily depends on the recommendation algorithm.

3.2 Graph-based approaches

Perugini et al. argue in (Perugini, Gonçalves, and Fox 2004), that all recommender systems make connections among people. These connections are either made directly - as a result of explicit user modeling, or indirectly - through the discovery of implicit relationships. This perspective is reflected in the notion of representing on-line, social interactions of users as a **social graph**, and their interests as their **interest graph**. Companies offering products that involve personalization recognize the benefits of graphs and employ graph-based algorithms, such as EdgeRank[1] used in Facebook's[2] activity stream recommendations.

In order to exploit the interest graph for the purpose of personalisation, an algorithm that can operate on a graph-based data structure is required. Next, we dscuss two graph-based personalisation approaches: (i) **Semantic distance** that provides a distance metric between entities of a semantic graph based on the number of direct and indirect links between two entities and (ii) **Spreading activation** that provides an iterative algorithm for identifying a set of related entities in a semantic graph.

Semantic distance Rada et al. (Rada et al. 1989) introduced the notion of using the distance between two entities in a semantic network as a proxy for conceptual distance. They define their conceptual distance as the minimum number of edges separating two nodes *a* and *b*. Passant (Passant 2010b) builds on the work of Rada et al. by defining a measure for the semantic distance between Linked Data entities - Linked Data Semantic Distance (LDSD). Passant proposes different combinations of direct and indirect links between entities, as well as weighted versions that give more weight to less popular links. This is motivated by the fact that two resources are more related if they are the only ones sharing a particular property. The different measures are then evaluated by users for the task of providing recommendations. A user based evaluation shows that the weighted semantic distance that takes both direct and indirect links into account performed best. (Passant 2010a) then shows that the this semantic distance measure can be used as the foundation for a music recommender system which uses DBpedia data. In addition, the different steps for calculating the semantic distance can be used to explain to the user why a certain artist was recommended (e.g. because two musicians use the same instrument or because they are on the same record label).

[1]https://www.facebook.com/note.php?note˙id=206484249362078

[2]http://facebook.com

The major drawback of the Linked Data Semantic Distance lies in lack of semantics of link types in the distance metric. However, existing research indicates that it not only provides a good measurement of similarity, but also enables finding interesting entities related to what the user already knows, but that are at the same time diverse. These properties make Linked Data Semantic Distance a valuable baseline for comparisons with other, more sophisticated graph-based recommendation approaches.

Spreading activation Spreading activation is based on previous work on semantic memory and case semantics (Cohen and Kjeldsen 1987). It is inspired by the fact that human memory retrieves memories by association. The idea was first implemented in the form of associative retrieval in database systems and was further developed out of the associative retrieval. Crestani (Crestani 1997) explains that the spreading activation requires a semantic network with directed (often typed or weighted) edges as data structure. If, as in the case of RDF, only typed edges exist in the graph, then different types can be associated with different weights. Two variations of network processing algorithms are recognized: **basic** spreading activation or **constrained** spreading activation.

Basic spreading activation: The inputs of the algorithm include the nodes that are activated at the start of the algorithm, which can represent the query or the interests in a user profile. With each pulse the activation spreads through the network. After a number of pulses (or after every single pulse) the termination condition is checked. The activated nodes after reaching the termination condition represent the most similar nodes to the initial activated set of nodes. Each iteration consists of two steps: (1) one or more pulses, which are propagated to all nodes which are directly connected to an active node, (2) a check of the termination condition, usually an upper limit to the number of activated nodes. Each pulse in turn is made up of three phases: (A) pre-adjustment, (B) spreading, and (C) post-adjustment. Each pulse of the spreading phase can activate new nodes. The spreading of the activation is determined by three functions of a node: the input function, the activation function and the output function.

The total input of a node is determined by the **input function**: $I_j = \sum_i O_k w_{ij}$, where I_j is the total input of node j; O_i is the output of node i connected to node j; and w_{ij} is the weight associated to the link which connects node i and node j. Activation along some edged (or edge types) can be controlled (suppressed or strengthened) through the use of weights. On the other hand, the **activation function** of a node is typically expressed as a threshold that is usually the same for each node. If the input of the node is higher then the threshold, then the node is active in the next pulse of the spreading phase. The **output function** determines how much output the node passes to all directly connected nodes in the next pulse. The output function can either produce a real value, which usually is the level of activation. Or the output function emits a binary value depending on its activation.

Constrained spreading activation: There are a number of considerable drawbacks of the basic spreading activation approach that include quick and simultaneous activation of all nodes in small and medium sized semantic networks or difficulties in exploiting the semantics of edges beyond the use of weights. Therefore, a number of constraints to limit the spreading activation process have been suggested (Crestani 1997). The **distance constraint** limits the activation to n links away from the original activated nodes, as in many scenarios the distance between entities relates to their similarity. The **fan-out constraint** poses that nodes with a high number of out-links should limit the spread the activation as they have a very broad semantic meaning, which can lead to the activation of nodes with a weak connection to the initially activated nodes. The **path constraint** relates to the semantics of the links, where types some links are activated as they are relevant to the use case or recommendation scenario. The **activation constraint** poses to control the total level of activation across the network to reflect different levels of importance of given nodes. These constraints can be seen as acting during the pre- and post-adjustment phases.

Relation to other approaches The spreading activation algorithm is superficially similar to the activation of **simulated neural networks**. However in a neural network simulation, the individual nodes do not represent semantic concepts, and the links between the entities do not represent any kind of semantic or conceptual connection or association between the nodes. A neural network is trained to reach a certain level of output activation when it is presented with a given input. The activation functions and the weights between the nodes are manipulated during the training phase in order to achieve this result. In contrast, each node of the semantic network which is used for spreading activation represents a certain conceptual entity, and the links only exist because they also exist in the domain on which the semantic network is modelled. The weights are not trained but they are based on the semantics of the modelled domain.

Another approach which is similar to spreading activation is the **PageRank algorithm**, or any other form of global, **eigenvektor-based algorithm**. As Berthold et al. argue in (Berthold et al. 2009), the continuous execution of the spreading activation algorithm without checking of the termination function will converge in a resulting activation state which can be interpreted similar to the result of the PageRank algorithm. The explanation for this phenomenon is based on the fact that the random graph walker which is employed to determine the PageRank of all nodes in the graph, basically executes a continuous and unconstrained form of spreading activation.

Most existing research on spreading activation is in fact using spreading activation on graphs with only weighted and no semantic edges. Related work includes Cohen et al. (Cohen and Kjeldsen 1987), Griffith et al. (Griffith, Oriordan, and Sorensen 2006) and (Kleb and Abecker 2010). The lack of the exploitation of the edge semantics can be attributed to various reasons. Cohen et al. (Cohen and Kjeldsen 1987) did not have the computational means to process a large semantic network and therefore used a more simple weighted graph for their spreading activation experiment. Griffith et al (Griffith, Oriordan, and Sorensen 2006) did not use semantic links between the nodes which represent users and items. Constrainted SA on numerical weights is then in fact used as a computationally less expensive alternative to existing collaborative filtering algorithms. Finally Kleb et al. (Kleb and Abecker 2010) use SA on a weighted graph in order to find relationships between similar entities.

The research on SA using weighted graphs shows that it is important to take the semantics of the network into account in order to exploit the full potential of the spreading activation approach, which differentiate it from other approaches such as neural networks and eigenvektor-based methods.

4 Application of personalisation approaches in an enterprise use case

Application of the spreading activation in the discussed use case (see Section 1) requires the ability to connect users and recommendable items into one unified semantic network. This enables the execution of the previously described spreading activation algorithm on semantic networks with three types of entities, like the example in figure 2.

The first type of entities represent **users** and their connections to communities (e.g. the user Conan). Second, there are entities that represent recommendable **items**, that is, social content such as blog posts or wikipages. Finally, there are entities which represent abstract **concepts**, typically used in tagging, often from external sources, linked with typed and directed relations.

Entities representing users are stored in user databases of all the systems for which content needs to be personalised. The nodes that represent recommendable social media are stored in all systems that contain user generated content (i.e. enterprise social platforms). The semantic network is augmented by linked data of DBPedia. DBPedia provides an RDF-based semantic network of all concepts from Wikipedia and their relations, which provide instance data as well as a background ontology. The final part of this example is formed by the links between users and the semantic network, and between the social media content and the semantic network. These links are enabled by the tags which users have used to describe their interests and expertise, and by the tags which users have assigned to the user generated content.

In terms of the data requirements for spreading activation, this translations to the following: The SA requires that the **user model** is provided by a list of concepts (mapped to a DBPedia URI) explaining user's interests/expertise, so that the user entity is connected to the semantic network. The **background data** is provided by the user generated content and the DBPedia RDF dataset. The spreading activation **algorithm** can be executed on the whole semantic graph with the start of activation in the user nodes. The termination condition is reached when a certain number of user generated content nodes is activated.

To avoid the cold-start and the empty-box problems, users and items must be connected to the DBPedia semantic net-

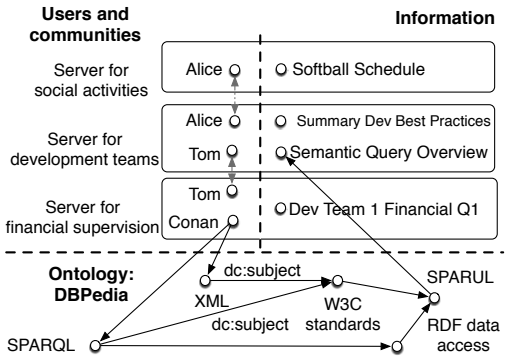

Figure 2: Spreading activation example with data sources

work. If this condition is met, the SA algorithm can provide generic content recommendations, which also enables the use of data from multiple sources (see Figure 2). As DBPedia provides a cross-domain and cross-genre semantic network, the SA algorithm can provide recommendations from different domains. Customisation for specific domains is possible via constraining of activation based on the link type. The spreading activation can also recommended similar users as sources, if the activation of user entities is considered in the termination condition. Scalability is achieved through the limitation of activation spread to only a few pulses. The algorithm does not depend on global properties such as the total number of links between nodes, but on local properties such as density of outward links from activated nodes. In addition, most of the activation states can be pre-computed depending on the starting activation node with the final global activation state computed at runtime.

5 Conclusions

In this paper we have reviewed the personalisation approaches that are applicable in enterprise social networks, which often require the use of data from different data sources and from different content domains. We first presented an enterprise use case and identified requirements for personalization methods. We argued the importance of the following requirements: mitigation of the cold-start and empty-box problems, scalability, applicability to multi-source and cross-domain data, universality regarding the data type and format, the ability to recommend sources of items, and the ability to customise generic recommendations for a specific domain. Further, we explained how the Spreading Activation approach can be applied in the enterprise use case to exploit the multi-source, cross-domain data available in enterprise social networks.

Our future work includes both qualitative and quantitive evaluation of the Spreading Activation. We plan to perform a qualitative comparative evaluation with the state-of-the-art methods, such as Collaborative Filtering (CF) or content-based methods. We will also perform a set of quantitative experiments and user studies to evaluate performance of the proposed method both in using standard Information Retrieval measures (e.g. precision/recall) and user perceptions.

References

Amini, B.; Ibrahim, R.; and Othman, M. 2011. Discovering the impact of knowledge in recommender systems: A comparative study. *International Journal of Computer Science and Engineering Survey (IJCSES)* 2(3):1–14.

Berthold, M.; Brandes, U.; Kötter, T.; Mader, M.; Nagel, U.; and Thiel, K. 2009. Pure spreading activation is pointless. In *Conf. on Information and knowledge management*.

Burke, R. 2002. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction* 12(4):331–370.

Cohen, P., and Kjeldsen, R. 1987. Information retrieval by constrained spreading activation in semantic networks. *Information processing & management* 23(4):255–268.

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

Deshpande, M., and Karypis, G. 2004. Item-based top-n recommendation algorithms. *ACM Trans. on Information Systems (TOIS)* 22(1):143–177.

Griffith, J.; Oriordan, C.; and Sorensen, H. 2006. A constrained spreading activation approach to collaborative filtering. In *Knowledge-Based Intelligent Information and Engineering Systems*, 766–773. Springer.

Herlocker, J.; Konstan, J.; Terveen, L.; and Riedl, J. 2004. Evaluating collaborative filtering recommender systems. *ACM Trans. on Information Systems (TOIS)* 22(1):53.

Ilyas, I.; Beskales, G.; and Soliman, M. 2008. A survey of top-k query processing techniques in relational database systems. *ACM Computing Surveys (CSUR)* 40(4):11.

Kleb, J., and Abecker, A. 2010. Entity reference resolution via spreading activation on rdf-graphs. *The Semantic Web: Research and Applications* 152–166.

Montaner, M.; López, B.; and De La Rosa, J. 2003. A taxonomy of recommender agents on the internet. *Artificial intelligence review* 19(4):285–330.

Passant, A. 2010a. dbrec – music recommendations using dbpedia. *The Semantic Web–ISWC 2010* 209–224.

Passant, A. 2010b. Measuring semantic distance on linking data and using it for resources recommendations. In *Linked AI: AAAI Spring Symposium" Linked Data Meets Artificial Intelligence", AIII*.

Pazzani, M., and Billsus, D. 2007. Content-based recommendation systems. *The adaptive web* 325–341.

Perugini, S.; Gonçalves, M.; and Fox, E. 2004. Recommender systems research: A connection-centric survey. *Journal of Intelligent Information Systems* 23(2):107–143.

Rada, R.; Mili, H.; Bicknell, E.; and Blettner, M. 1989. Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on* 19(1):17–30.

Sarwar, B.; Karypis, G.; Konstan, J.; and Reidl, J. 2001. Item-based collaborative filtering recommendation algorithms. In *International Conference on the World Wide Web*, 285–295.