Matching Resources in Social Environment

Amel Benna[1,2], Hakima Mellah[1], Islam Choui[3] and Ali Oualid[3]

[1]CERIST, 05, Rue des 03 Frères Aissiou, BP 143, BenAknoun, 16030 Algiers, Algeria

[2]USTHB, BP 32, El-Alia Bab-Ezzouar, Algiers, 16111, Algeria

[3]ESI, BP 68M, Oued Smar, 16309, Algiers, Algeria

Abstract. User comments on the web are becoming more and more important. We focus, in this paper, on the use of user-defined tags for annotating resources to identify links between them. These links are based on a social context of the resource, obtained by applying k-means classification method and a hierarchical classification of tags within a cluster. The resources are re-assigned to this classification to facilitate the search process. The ranking of results is performed according to their degree of relevance, by evaluating a similarity score between the tagged contents, in hierarchical clusters of tags, and the user request. The results of the evaluation, on the social bookmarking system del.icio.us, demonstrate significant improvements over traditional approaches.

1 Introduction

User experience and comments on the web are becoming more and more important. In 2010, Gartner group, predict that within five years, 70 percent of collaboration and communications applications designed on PCs will be modelled after user experience lessons from smart-phone collaboration applications[4].

A collaborative tagging system put users at the centre of data production and introduces a strong social collaboration. It describes the process by which many users add meta-data in the form of keywords to shared contents [1]. These keywords require no skill from user and are named tags. They and can be associated with different types of resources (videos, images, bookmarks, articles, application and blogs).

The analysis of collaborative tagging systems structure showed regularities in user activity, tag frequencies, kinds of tags used and a remarkable stability in the relative proportions of tags within a given resource. Empirically, once a resource has been tagged over a hundred times, each tag's frequency, in a proportion, remains stable compared to the total frequency of all other tags used for this resource [1]. However, works on linkage information often do not take into account social information of resource that can be retrieving from users significant tags. Indeed, the Social Information Retrieval (SIR) follows from its domain model [2] and the incorporation of social factors can increase the relevance of results returned in information retrieval [3],[4],[5],[6].

[4] http://info.absnt.com/

Benna A., Mellah H., Choui I. and Oualid A..
Matching Resources in Social Environment.
DOI: 10.5220/0004088800610070
In *Proceedings of the 10th International Workshop on Modelling, Simulation, Verification and Validation of Enterprise Information Systems and 1st International Workshop on Web Intelligence* (WEBI-2012), pages 61-70
ISBN: 978-989-8565-14-3

Our interest is focused on the use of user-defined tags for annotating resources to identify links between resources. These links are based on a social context of the resource in folksonomy. A folksonomy is a system of classification derived from the practice and method of collaboratively creating and managing tags to annotate and categorize contents [7]. A resource social context is related to purified and classified tags, obtained by a classification method, and the words content refers to a resource or portion of resource.

More specifically, we propose an approach for SIR that integrates social relationships between contents by taking into account the social information of resource. A social linking between resources is based on clusters. These latter are, a set of semantics links between users purified tags, obtained by a classification method. We operate the most significant tags in research. E.g. The adjectives tags such as *"funny"*, *"interesting"*, *"mydocument"* or words they do not even exist in the literature as *"xfd4"* are eliminated for low frequency use. We reduce redundancy or ambiguity of tags by finding semantically related tags as tags from the same cluster. The proposed process for modelling links between resources classifies and structures a folksonomy and includes it in the matching and the ranking of search results.

The rest of this paper is structured as follows: Section 2 introduces some related works on SIR based on linkage information and collaborative tagging systems. Section 3 proposes our model for linking resources in information retrieval based on collaborative tagging. Section 4 presents the results of our evaluation. We conclude and highlight future research direction in section 5.

2 Related Work

Several approaches have been proposed for using linkage information on the Information Retrieval (IR) systems. This work can be distinguished according to different factors (content, HTML, architecture, links, social, trust, personal)[5].

The most famous proposed approaches for using linkage information, to aid in relevant document retrieval, are PageRank [9] and HITS [10]. PageRank technique is calculated independent of any query but pages with high PageRank are highly ranked even though they are not relevant to a user's query. Unlike PageRank, HITS is a query-dependent form of linkage analysis. Two scores; authority and Hub are calculated for each document. Evaluation of model [11], based on the works [2],[12], proves that the extent of Hub (the centrality of the authors) is the measure to better assess the social significance of documents. However, if the initial query expressed by a user does not cover a sufficiently broad topic, there will often not be enough relevant pages. The main disadvantage of this approach is that not only requires extra resources from the search system at query time but also increases the system response time. The model proposed in [17] illustrates an example of a study by applying four centrality measures (degree, PageRank, closeness and betweenness) to evolving co-authorship network. In this work, the measures of centrality include the impact of the resource, i.e. its citing accounts and scope of author.

[5] From Periodic Table of Search Engine Optimisation Ranking Factors (http://SELND.COM/SEOTABLE)

Search engines such as *Google, Yahoo and Bing* use several factors to retrieve information, some factors may influence more than others and may be considered more important than others. However, no single factor guarantees a relevant research and top rankings[4].

To improve the web search, various approaches [4],[5],[13] explore the use of social annotations. In [4] two new algorithms are proposed: the first one calculates the similarity between social annotations and web queries whereas the second captures the popularity of web pages using social annotations. A model in [5], based on social approval votes of documents, shows that social information on documents can improve research and the approval sources provide more details on user needs, particularly, when votes are provided by experts. To define user expertise level, a user model in [13] is integrated in calculating the tag weight. The evaluation is based on the closeness degree between user interest's and resource area, expertise and personal assessment for tags associated to the resource.

Nevertheless, IR systems that use collaborative tagging suffer from a number of limitations such as: variability on writing some tags, ambiguity due to the existence of synonyms, an the absence of semantic links between tags. These leads to impoverish information research potential whereas the rate of tagged contents is growing every day, and affect the response time and the result quality. Data clustering has been used, for natural classification, to identify the degree of similarity among forms or organisms, and for compression, as a method for organizing the data and summarizing it through cluster prototypes. A cluster of tags represents the most common way to gather additional information in collaborative tagging systems [8]. It was defined to:
- Use the most significant tags [14],
- Decrease redundancy or tags ambiguity [15],
- Find the similar semantic tags [14],
- Reduce the response time and improve the quality of results[16]. Thousands of clustering algorithms have been proposed in the literature. Nevertheless, clustering methods differ on the choice of the objective function, probabilistic generative models, and heuristics [8]. The K-means [8] method is used to classify tags of folksonomies such as: customizing folksonomies based clustering of tags [14], extraction of relationships between users and resources tagged based clusters of tags [18]. The experience on Wordnet ontology, in [19], showed that the tags associated through simple co-occurrence measures tend to maintain subsumption relationships (a hierarchical relationship between concepts), whereas tags associated via a similarity distributional measure in the context tag-tag tend to be at the same hierarchical level, or to share the same parent/ grandparent.

The works on linkage information we have cited do not take into account social information of resources, which can be retrieved from significant users tags; The social factor focuses only on a social reputation of a user account and on user social shares in social network and neglect social information of resources and links between tags. Inspired by this works and to consider social factors for using linkage information, we modelled links between resources for social research, using a purified tags and a hierarchical structure of cluster of tags.

3 Social Linking Model

The social linking model, based on practices of collaborative tagging, is used to define links between resources (see Fig. 1). The definition and the structure of the proposed folksonomy for linking resources are presented in section 3.1. We describe the social search process that explores folksonomy in section 3.2 and the evaluation of results in section 3.3.

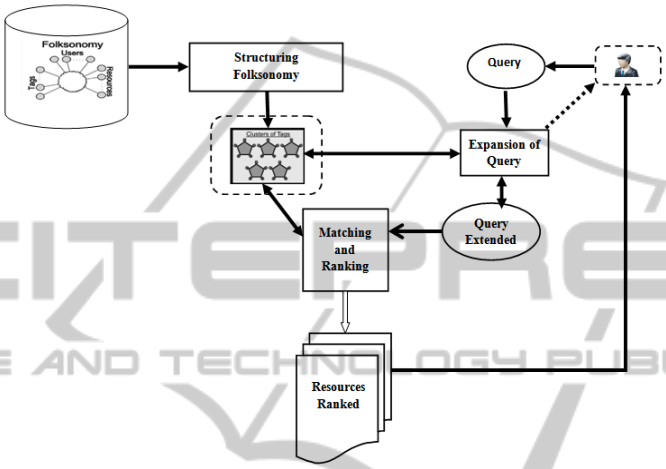

Fig. 1. Social linking resources architecture.

3.1 Folksonomy Structure

Let consider *F=(U,T,R,Y)* the formal structure of a folksonomy [22]. *U*, *T* and *R* are finite sets, whose elements are respectively users, tags and resources. *Y* is a ternary relation between them such that:

$$Y \subset U \times T \times R$$

A post is a triple (u, t_{u_r}, r), where, $t_{u_r} \in T$ is a tag used by user $u, u \in U$ to tag a resource $r, r \in R$. The classification process of our folksonomy includes four steps:
1. Creating a semantic tag-tag data matrix,
2. Generating clusters of tags,
3. Defining hierarchy of tags within a cluster,
4. Assigning resources to cluster of tags hierarchy.

Semantic Tag-tag Data Matrix. To define a link between tags we compute the co-occurrence matrix in the context tag-tag [19]. This co-occurrence is determined by the co-occurrence $W(t_i, t_j)$ between each pair of tags (t_i, t_j) as in (1).

$$W(t_i, t_j) = |(u, r) \in U \times R/(u, t_i, r) \in Y(u, t_j, r) \in Y| \quad (1)$$

The data matrix is then transformed into a cosine matrix [20] by measuring the cosine distance between vectors, as in (2), where a vector, $\overrightarrow{v_{t_i}}$, denotes the number of times a user U_i uses a tag t_j and it is computed as described in [23].

$$\cos(\overrightarrow{v_{t_i}}, \overrightarrow{v_{t_j}}) = \frac{\overrightarrow{v_{t_i}} \cdot \overrightarrow{v_{t_j}}}{\|\overrightarrow{v_{t_i}}\| \cdot \|\overrightarrow{v_{t_j}}\|} \tag{2}$$

Generating Clusters of Tags. To minimizing within clusters variance tags and maximizing the distance between clusters of tags, we apply the k-means method on the cosine matrix. The K-means algorithm requires three user-specified parameters: number of clusters K, cluster initialization, and distance metric. The most critical choice is K. Whereas no perfect mathematical criterion exists; a number of heuristics are available for choosing K. One way to overcome the local minima is to run the K-means algorithm, for a given K, with multiple different initial partitions and choose the partition with the smallest squared error.[8] After applying k-means to generate set of k-clusters of tags, we use the Levenshtein distance [21] to avoid spelling variations of tags and composite words in each cluster of tags.

Cluster of Tags Hierarchy. A hierarchy of tags in each cluster is build. Each tag, with its variant spellings are grouped into a single concept by applying hierarchical classification algorithm [22]. This hierarchy structures the clusters of tags as a tree, where tags are tree nodes and resources tree leafs. We design by tags path, any path leads from the root node (the most common tag used in the cluster) to a leaf node (tags used less in the cluster). The tag that has a high degree of co-occurrence in the resources is chosen as a concept.

Assigning the Resources to Cluster of Tags Hierarchy. In order to form clusters that contain similar resources, the resources tagged are reassigned first to clusters of tags. A resource, r_i, degree of membership, $D_{r_i c_j}$, to the cluster c_j is computed as in (3). $occ(t_l, r_i)$ denotes co-occurrence of tag t_l with a resource r_i and t_l belongs to cluster c_j.

$$D_{r_i c_j} = \frac{\sum_{t_l \in c_j} occ(t_l, r_i)}{|u \in U/(u, t, r_i \in Y)|} \tag{3}$$

To determine resource tags in cluster of tags, each resource, r_i, is associated to the hierarchical cluster of tags whose degree of belonging to it is maximal.

After having classified the folksonomy F into clusters of tags, defining hierarchy of tags in each cluster, and reassigning resources to tags, an XML file is used to store the structuring folksonomy, i.e. hierarchy of clusters tags, tags and associated resources.

3.2 Social Search Process

To answer a user query, the first step of the social search process is the query expansion, the second one is the matching between clusters of tags and request tags and the last step is the ranking of results.

Query Expansion. When a user issues a query, it is disambiguated by detecting variations in spelling of its keywords. The Levenshtein distance is used with a threshold equal to 0.8. Indeed, most tags are names, and thus the lemmatization methods are not recommended [15]. After query disambiguation, a linguistic ontology is used to determine semantic of request tags. In fact, request keywords are considered as tags. The objective is to guide the user by suggesting keywords related to the meaning of the request word.
E.g. When the WordNet ontology is used for the word Java, three senses are proposed: island, coffee, object-oriented programming language. The user request is enriched by tags *language*, *object- oriented*, and *programming* for *computer science* user interest area.

Matching and Ranking Resources. To answer a user find resources, we first identify clusters of tags that match with the user request tags, and then search for tagged content matching user query. Because tags are structured in hierarchical cluster, the user query tags can match the cluster of tags tree nodes. As tags are close together, there is great probability that request tags belong to the same cluster of tags. To identify clusters of tags matching user request, a semantic similarity score, $Jaccard(\overrightarrow{V_r}, \overrightarrow{V_{c_i}})$, is computed between each vectors of clusters, $\overrightarrow{V_r}$, and query expansion vector, $\overrightarrow{V_{c_i}}$, as in (5). The user request is represented by a tags vector,$\overrightarrow{V_r}$, and each cluster of tags, c_i, is represented by tag vector $\overrightarrow{V_{c_i}}$.

$$Jaccard(\overrightarrow{V_r}, \overrightarrow{V_{c_i}}) = \frac{\overrightarrow{V_r} \sqcap \overrightarrow{V_{c_i}}}{\|\overrightarrow{V_r} \sqcup \overrightarrow{V_{c_i}}\|} \tag{4}$$

After identifying cluster of tags, identifying resources that meet user request means to browse tree, of the selected clusters of tags, looking for tree leaves in which nodes match the query tags. To determine such leaves, for each cluster of tags whose tags match the tags of the query we proceed as follows:
1. For each node, we select all the related contents where nodes tags match user request tags.
2. If in the same path, there is more than one tag that matches user request tags, we select the deepest one in the subtree.

E.g. let *language*, *object*, *java*, *javascript* be tags of user expansion request for query keyword *Java*. The tags *Object* and *Java* are in the same tree path (see fig. 2), but the tag *Java* is deeper than the tag *Object* in the tags hierarchy. The content C_{T_3} tagged by *jsp* is select as request result. The contents C_{T_1}, C_{T_2} are also selected for the *javascript* tag, as leaf of node *javascript* in cluster of tags subtree.

Results Ranking. The responses to a query may be found in a single content or may be subject to an aggregation of a set of results shared with different contents. The aggregation of contents, returned by query results, is to combine contents that match the user request but that was tagged by different users of the system. This aggregation includes any type of resource (text document, image and video). For the example in Fig. 2 the contents C_{T_1}, C_{T_2} *and* C_{T_3} are aggregated and are displayed to end user as one result. The ranking of resources returned in a search is performed according to their degree of

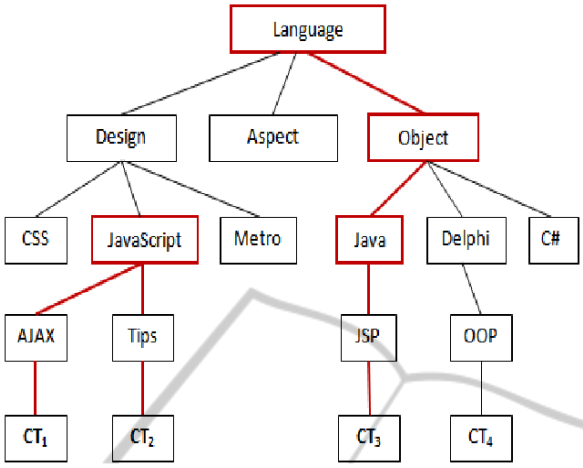

Fig. 2. Example of hierarchy of tags

relevance. We compute similarity score, $Sim(\overrightarrow{V_{t_{c_i}}}, \overrightarrow{V_r})$, between request and content by using the, most commonly used measure, cosine of the angle between the query vector, and tagged content vectors. This score is computed as in (6). T_k denotes the k^{th} tags of request.

$$Sim(\overrightarrow{V_{t_{c_i}}}, \overrightarrow{V_r}) = \frac{\sum_{t_k \in \overrightarrow{V_r}} occ(t_k, \overrightarrow{V_{t_{c_i}}})}{|u \in U/(u, t, r_i \in Y)|} \quad (5)$$

3.3 Evaluation and Results

To evaluated our approach, we extracted data from *'delicious data'* which contains a set *U* of 2000 users, a set *T* of 2000 tags, and a set *R* of 70 resources for 3577 annotations. Data analysis for 2000 tags showed that 200 tags have a high co-occurrence frequency for 1879 users and represent more than 70% of users annotations. We had used only the triplets of tags, users and resources. These triplets represent 80% of folksonomy tags.

As a first step, we seek to build clusters of tags (see Fig. 3). The set *T* of tags has been classified using *k-means* method, with *k = 17*. The result is a set *C* of clusters with an average of 12.3 tags for each cluster. The similarity distance between two resources assigned to a cluster is greater than 0.65. The resources that have a similarity degree more than 0.9 are grouped in the same slice of branch in the hierarchy To Define a hierarchy of the ordered list in cluster of tags,we measure the cosine similarity between vectors for $\lambda > 0.5$.

To evaluate the relevance of our approach, a series of tests for two kind of research are performed: a traditional IR, based on the vector model, and a SIR, based on the model that we define in section 3. Fig. 4 illustrates an example of the relevance measure, for the top 5 recommended resources, for a query *Java* in the *music* interest area. The recall-precision curves measures vary inversely, precision decreases as the recall increases. We observed that SIR search performs better than traditional IR search.

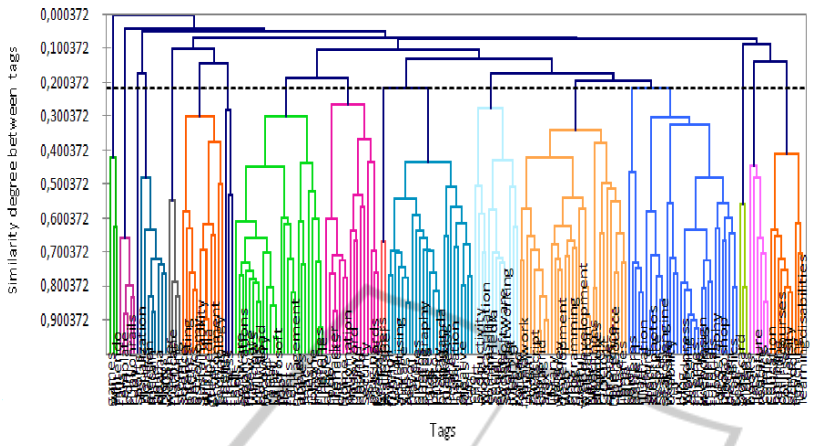

Fig. 3. Dendrogram of similarity between tags.

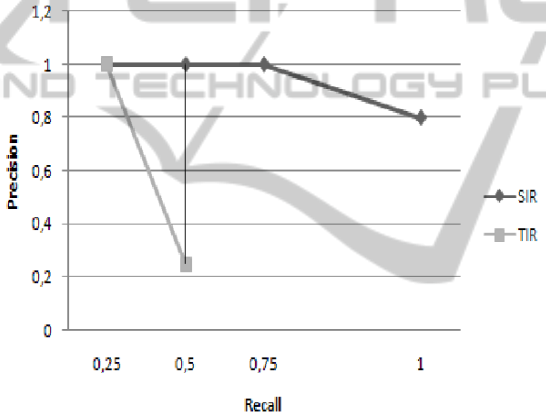

Fig. 4. Example of recall and precision evaluation for query.

4 Conclusions

We describe the social linking model based on practices of collaborative tagging to define links between resources. The structure of the proposed folksonomy allows assigning resources to hierarchical structure of tags within cluster. We apply the clustering algorithm, k-means, to define clusters of tags and hierarchy in each cluster. A user query can be expanded, by suggesting tags from the cluster of tags and ontology. In our social search process we define a similarity degree for ranking function to classify aggregated contents according to their relevance degree.

The evaluation was flown on the bookmarking system del.icio.us. The first conclusions that emerge from evaluation of relevance tests are that the similarity distance between resources within cluster of tags are very close and the integration of resource social context provides very conclusive over traditional approaches. Our future works are ori-

ented towards the definition of local ontology from the hierarchy of tags within clusters of tags.

References

1. Golder, S., Huberman, A. B.: The Structure of Collaborative Tagging Systems, CoRR abs/cs/0508082. 18 August 2005.
2. Kirsch, S.M., Melanie, G., Cremers B. A.: Beyond the Web: Retrieval in Social Information Spaces,.In Advances in Information Retrieval. London, UK : Springer, Lecture Notes in Computer Science,2006, Vol. 3936, 84-95.
3. Zanardi,V., Capra,L.: Social Ranking: Uncovering Relevant Content Using Tag-based Recommender Systems, RecSys'08, 23-25 October 2008.
4. Bao, S. , Gui-Rong, X., Xiaoyuan,W., Yong, Y., Fei, B., Su, Z.: Optimizing web search using social annotations, In Proceedings of the 16th Internationa Conference on World Wide Web, WWW 2007. ACM 2007, 8-12 May 2007, 501-510.
5. Kazai, G., Milic-Frayling, N.: Effects of Social Approval Votes on Search Performance, In ITNG 2009, Sixth International Conference on Information Technology: New Generations, 27-29 April 2009, ISBN 978-0-7695-3596-8, 1554-1559.
6. Benna, A., Mellah, H., Hadjari, K.: Building a social network, based on collaborative tagging, to enhance social information retrieval, In ICITES, 2012, 453-458.
7. Isabella, P.: Folksonomies. Indexing and Retrieval in Web 2.0, K G Saur Verlag, 2009.
8. Jain, A.:Data Clustering: 50 Years Beyond K-Means, 2009.In Pattern Recognition Letters, 2009.
9. Page, L., Brin, S., Motwani, R., Winograd, T.:The PageRank citation ranking: Bringing order to the web, In WWW98. 1998, 161172.
10. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment,In SODA, ACM, 1999, Vol.46, 668-677.
11. Ben Jabeur, L., Tamine, L., Boughanem, M.: A social model for literature access: towards a weighted social network of authors,RIAO 2010, 32-39.
12. Mutschke,P.: Enhancing Information Retrieval in Federated Bibliographic Data Sources Using Author Network Based Stratagems, In ECDL 2001, LNCS 2163, pp. Springer, 4-9 September, Vol. 2163, 287-299.
13. Kichou, S. , Mellah, H., Amghar, Y., Dahak F.: Tags Weighting Based on User Profile, In Active Media Technology - 7th International Conference, AMT 2011. Lecture Notes in Computer Science 6890 Springer 2011, 7-9 September 2011, ISBN 978-3-642-23619-8, 206-216.
14. Gemmell, J., Shepitsen, A., Mobasher, B., Burke, R.D.: Personalizing Navigation in Folksonomies Using Hierarchical Tag Clustering, In DaWaK 2008, 196-205.
15. Spiteri, L.: Structure and form of folksonomy tags: The road to the public library catalogue, 2007, Vol. 4.
16. Begelman,G., Keller P., Smadja, F.: Automated tag clustering: Improving search and exploration in the tag space, Proc. of the Collaborative Web Tagging Workshop at WWW., May 2006, 2226.
17. Yan, E., Ding, Y.: Applying centrality measures to impact analysis: A coauthorship network analysis, 2009.
18. Grahl, M., Hotho, A., Stumme, G. : Conceptual Clustering of Social Bookmarking Sites, In LWA. Workshop. September 2007, 50-54.
19. Cattuto, C., Ben, D., Hotho, A., Stumme, G. : Semantic Grounding of Tag Relatedness in Social Bookmarking Systems, In The Semantic Web - ISWC 2008, 7th International Semantic Web Conference, 26-30 october, 2008, 615-631.

20. Cattuto, C., Loreto, V., Pietronero, L. : Collaborative Tagging and Semiotic Dynamics,In CoRR. May 2006.
21. Levenshtein, V. : Binary codes capable of correcting deletions, insertions, 1966.
22. Hsieh, W-T., Lai, W-S., Chou, S-C. : A collaborative tagging system for learning resources sharing, In IV International Conference on Multimedia and Information and Communication Technologies in Education (m-ICTE2006), 2006, 1364-1368.
23. Markines, B., Cattuto, C., Menczer, F., Benz,D., Hotho, A., Stumme, G.: Evaluating similarity measures for emergent semantics of social tagging. In WWW 2009, 2009, 641-650.