Project no. 34721

TAGora

Semiotic Dynamics in Online Social Communities

http://www.tagora-project.eu

Sixth Framework Programme (FP6)

Future and Emerging Technologies of the Information Society Technologies (IST-FET Priority)

D3.3 Methods for using semantic inference in data analysis

Period covered: from 01/06/2007 to 31/05/2008
Start date of project: June 1st, 2006
Due date of deliverable: July 15th, 2008
Distribution: Public

Date of preparation: 31/05/2008
Duration: 36 months
Actual submission date: June 20th, 2008
Status: Final

Project coordinator: Vittorio Loreto
Project coordinator organisation name: “Sapienza” Università di Roma
Lead contractor for these deliverables: UNIKO-LD

Executive Summary

This deliverable presents how to improve the data analysis with semantic inferencing by taking background knowledge into account. The first part is about data pre-processing, i.e. how to filter and clean up the datasets, for the later analysis. This is quite important since it can significantly influence the analysis results. A filtering architecture which uses different source of background knowledge.

In the second part, actual techniques for semantic inferencing are presented. This includes the Triple Play approach which uses Latent Semantic Indexing to find hidden concepts in the datasets. It is applied on so called smoothed vectors space models, i.e. recombinations of the typical relation matrices between folksonomy items to overcome the data sparseness problem.

Moreover, there are two classification algorithms working on image and audio data. The first one uses machine learning techniques to classify tags describing locations. The second one is about exploiting tag correlations to improve acoustic classifiers.

Finally, measures of tag and resource similarity were explored systematically and characterized semantically by grounding them in formal representation of knowledge (WordNet, DMOZ).

Contents

List of Figures

Chapter 1

Introduction

1.1 Semantic Inference

The analysis of folksnonomy data based on statistical methods is not able to tackle all aspects of data relations as it is based on many independency assumptions. In fact, we can observe that the tags in a folksnonomy often stand in a relationship with others tags. Examples are relations to tags with e. g. a homonymous or synomymous meaning or relationships to tags in another language. Such relationships can only be resolved with some background knowledge underlying the folksonomy.

Background knowledge can be used for different purposes. For example, the folksonomy datasets often need to be cleaned up before the data analysis can be performed. In this case background knowledge can help to identify the less important information that can be filtered out. In contrast to that, background knowledge can also be used to extend the dataset with additional information. This is also closely related to the data sparseness problem. The sparseness in the folksonomy datasets comes from the fact that users usually annotate resources within different contexts and with just a few tags out of the set of all applicable data annotations. Thus similar resources may not be recognized as such because different annotations were assigned. This effect is especially significant in narrow folksnonmies like Flickr where a tag can only be assigned once to a photo. In broad folksnonomies like del.icio.us the same tag can be assigned to the same resource multiple times by different users. Thus, additional information like tag popularity (i.e. tag frequency) on a per-resource basis is available.

A simple way to overcome the sparsity problem is to derive such missing information from co-occurrence patterns. For example, if a user almost always uses the tags “web2.0” and “ajax” together one could assume that in the few cases where he does not he simply forgot to add the second one. Although that assumption may be correct in some cases we can not be sure about the semantic relation of the tags maybe synonyms and homonyms are involved. At that point the background knowledge comes into play with the goal to better differentiate between the different semantic meanings. This is especially useful for annotations on different semantic levels like super and sub concept.

1.2 Background Knowledge

Background knowledge can be taken from many different sources and for different purposes. Generally, folksonomy data is tripartite containing relation between the set of users, tags, and resources. Although background knowledge can be applied to all three of them it is most commonly applied to tags.

Background knowledge about a resource is derived from an analysis of the resource itself. In case

of bookmarks, e. g. taken from del.icio.us, we can parse the referenced documents and extract the containing terms. For images typically low-level features like dominant color etc. are extracted and considered as additional image annotations. Background knowledge for users usually comprises the users membership in user groups or friends relations. If a user has multiple profiles in different folksonomies such additional information can also be very useful. However, if not explicitly given the identification of such cross-folksonomy profiles is complicated and the verification is almost impossible.

Following, we just concentrate on background knowledge used for tags and point out the different applications. Within Tagora, background knowledge is primarily used for pre-processing and filtering the datasets or to create subsets that are better suited for the analysis task.

Folksonomy users quite often tend to use words of their respective language to describe resources. However, the greater part of the used tags can not be found in any dictionary. They are word/number combinations, names of places/persons, newly made up words or even misspellings of proper words. Wordnet is a dictionary used to identify words of the English language and find respective singular/plural forms as well as synonyms. Beyond that, tags can also be analyzed lexically including stemming to reduce them to a common form.

Taxonomies like DMOZ (`http://www.dmoz.org/`) can be used to categorize tags and put them in a hierarchy. This is especially useful to identify tags with more abstract (high-level) or concrete (low-level) semantic meaning. Even more powerful is the use of ontologies which provide exact definitions of the relations between concepts. The latter idea has been used to characterize and compare several statistical measures of tag “similarity” (or “relatedness”) computed on the folksonomy. This was achieved by means of semantic grounding: measures of tag similarity (Cattuto et al., 2008a,b) were grounded in Wordnet (`http://wordnet.princeton.edu/`), and measures of resource similarity (Markines et al., 2008) were grounded in the Open Directory Project (DMOZ). Semantic grounding allowed us to provide a semantic characterization of statistical measures on the folksonomy, affording a more formal definition of what is as a “related” tag, as well as formal metrics for comparing the performance of measures aimed at detecting tags that bear a specific semantic relation to a given tag, for example synonym tags. Eventually, this led to two main results, presented in Ref. (Markines et al., 2008): 1) a clear definition of classes of folksonomy-based measures that involve different levels of projection and aggregation, symmetrically defined on both tags and resources (and in principle, users); 2) an empirical analysis of the computational scalability of the measures introduced, showing that by suitably choosing projection and aggregation schemes it is possibly to devise measures that are both semantically accurate and scalable.

Another useful tool is Google Search. It allows to get feedback on a term’s popularity by its number of returned search results. This is also useful for finding common word combinations or to identify technical terms that are not included in common dictionaries. Moreover, Google Search provides spelling suggestions for potentially misspelled words. Thus, it is possible to correct spelling mistakes in the tags. Otherwise such tags are just noise in the dataset that is not usable for the analysis or can even negatively influence it.

1.3 Structure of this Deliverable

In this deliverable we will first describe in chapter 2 some concrete approaches used for dataset treatment before doing the semantic analysis. It is divided in data filtering and data enrichment. The data filtering part gives an overview of the techniques using background knowledge to remove noise from the data sets. Data enrichment is adding more information to the existing datasets (with or without background knowledge), e.g. by introducing new relations between the data items. This is mainly done to overcome the data sparseness problem and can significantly improve the results of the semantic analysis. Chapter 3 continues with the description of algorithms used for semantic inferencing. This includes Triple Play, a method employing Latent Semantic Analysis with

different matrix combinations to identify important concepts hidden in the datasets. Thereafter, a approaches for classifying images and sound data is described.

Chapter 2

Dataset Treatment

2.1 Tag filtering

Tags are free text, and users can tag resources with any terms they wish to use. On the one hand, this total freedom simplifies the process and thus attracts users to contribute. It also avoids the problem of forcing users into using terms they do not feel apply, as opposed to enforcing the use of a set of terms. For these reasons, the lack of constraints seems essential. On the other hand, it generates various vocabulary problems, where tags can be too personalised, made of compound words, mix plural and singular terms, meaningless, synonymous, etc. This total lack of control is resulting in some sort of tagging chaos, thus obstructing search and analysis.

Thus, while it is possible to gather information from multiple folksonomy sites, such as Flickr or del.icio.us, inconsistency will lead to confusion and loss of information when tagging data is compared. For example, if a user has tagged photos from a recent holiday in New York with nyc, but also bookmarked relevant pages in del.icio.us with new_york, the correlation will be lost. In order to improve folksonomy data analysis and integration, tags have to be filtered to increase their compatibility.

Therefore, we developed a tag filtering architecture that makes use of external knowledge resources such as WordNet, Wikipedia, and Google search engine. The filtering process is a sequential execution where the output from one filtering step is used as input to the next. The output of the entire filtering process is a set of new tags that correspond to an agreed representation. The figure below shows how the tag filtering process works:

The tag filtering processes are (numbered as in the figure):

1. Syntactic filtering: This process filters out too-short or too-long tags, converting non-English characters to their base form, removing stop words, etc. Remaining tags are then validated against WordNet. If a remaining is found in WordNet then it will be added to the filtered set straight away, thus bypassing any further filtering steps.

2. Compound nouns and misspellings: Tags suspected of being misspelled or made of several words are passed to Google as a search query. The Google “did you mean” service can suggest alternative spellings for query terms, and break up compound terms. Remaining tags are verified with WordNet as in the previous step.

3. Wikipedia correlation: WordNet coverage is rather limited and in spite of its large scale, it falls short of describing a wide range of entities. To this end, we use Wikipedia to look for people names, acronyms, companies, contemporary terminologies, etc. For each remaining terms in the tag set, we try to identify a Wikipedia entry. At this stage, if a term cannot be correlated with Wikipedia then it will filtered out.

4. Morphologically similar: This step and that one that follows look for tags that can be merged

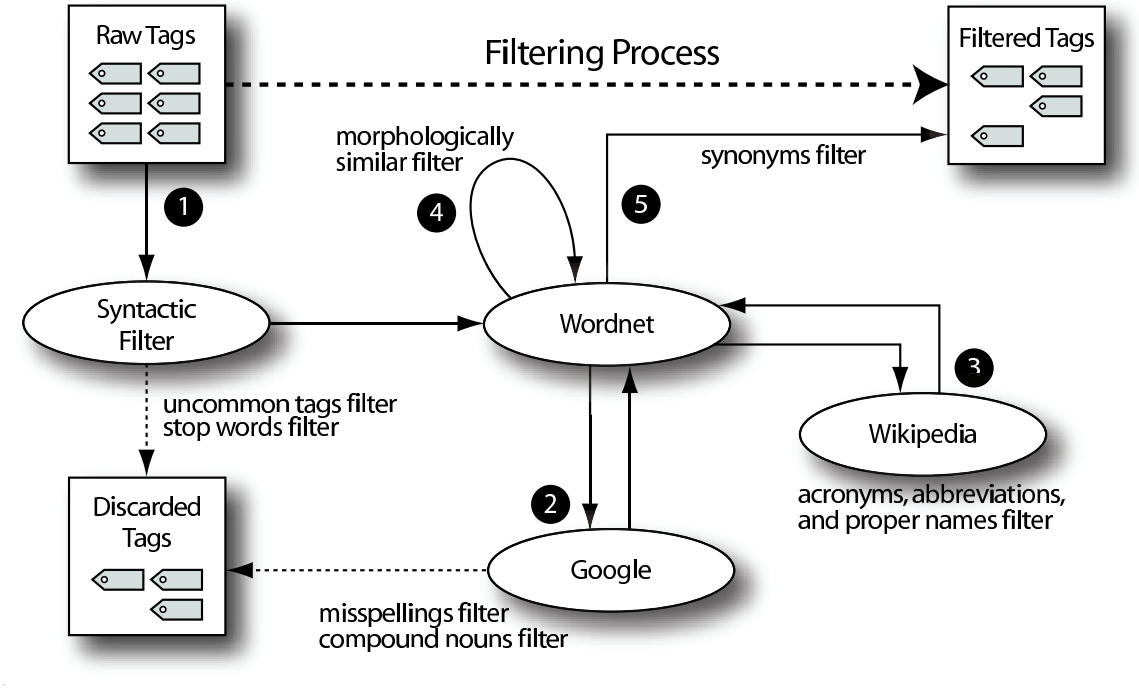

Figure 2.1: The tag filtering process

together. This step looks for, and merges, morphologically similar terms (e.g. blog, blogs, blogging).

5. WordNet synonyms: This last tag-filtering step uses WordNet to find synonymous tags and combine them accordingly.

We used the tag filtering process described above to clean user tags to (a) instantiate an ontology for multi domain recommendations (Cantador et al., 2008), (b) compare and match distributed user tag clouds to build user profiles (Szomszor et al., 2008a), and (c) measure the similarity of user tag clouds (Szomszor et al., 2008b).

2.2 Data Enrichment

Matrix representations of folksonomy relations are especially useful for semantic analysis. Typical relations matrices, as also presented in Deliverable 3.1. are the combinations of user-tag, user-resource, or tag-resource. Here, each row or column of the matrix stands for an element of the respective item set and the matrix entries define a weight for the relations between the corresponding items. Likewise, co-occurrence networks (e.g for tags) can be represented as matrices, too.

One advantage of these matrix representations is that they can be combined by typical matrix operations like multiplications. This is especially useful for tackling the data sparsity problem, which stems from the fact that the relation matrices are usually sparsely filled. By multiplying for example a user-tag matrix with a tag-resource matrix a much denser matrix is created that incorporates the information from the original matrices. This operation is among others applied in (Abbasi and Staab, 2008) to *smoothen* the dataset and then feed it to a Latent Semantic Analysis step for identifying hidden concepts that are hard to discover in the original relation matrices.

Chapter 3

Semantic Inference

3.1 Triple Play and Latent Semantic Analysis

A major problem in searching folksonomies is data sparsity. There are many resources which do not appear in search results because of sparseness of data. We introduce the method *Triple Play*, which smooths the tag space with the help of the user space. As a part of *Triple Play*, we proposed two new vector space models for folksonomies, *SmoothVSM Dense* and *SmoothVSM Sparse*. These vector space models exploit the user-tag co-occurrence relationship to overcome the problem of sparseness. Finally, we apply Latent Semantic Analysis to different vector space models and analyze the results. An initial evaluation shows that using the additional information that is available in folksonomies helps in improving search results. Figure 3.1 shows the overall process of *Triple Play*. Details of *Triple Play* can be found in the paper (Abbasi and Staab, 2008).

For the initial evaluation of Triple Play, we created the scenario of querying and retrieving resources. We assumed that usually the resources will not have assigned all relevant tags that may be used for retrieving them. It is the goal of TriplePlay to reproduce these missing tags with the help of background knowledge that is taken from additional information usually available in Folksonomies.

For our evaluation, we created an artificial data set by taking approx. 10,000 resources from a Flickr data set. From these resources, we removed some of the tags, i. e. we simulated the missing of tag information. The artificial data set was then used as the input for Triple Play and it was evaluated in how it was possible to also retrieve the resources with the missing tags. The evaluation showed that the Vector Space Models proposed in Triple Play produced better results than the simple Vector Space Models that do not exploit the additional information available in Folksonomies (see (Abbasi and Staab, 2008)). A detailed TREC based evaluation of Triple Play is currently being done.

3.2 Image Classification

When planning the next holiday people often use photo sharing services like Flickr for seeing pictures of representative landmarks in their vacation destination. Recently proposed geo-tagging based algorithms automatically construct such landmark photo summaries. However, these algorithms are not applicable to locations with a low number of geo-tagged photos. We propose a method to identify city landmarks using only the more common natural-language tags. For finding landmark photos we apply a SVM classifier trained with positive examples from Flickr groups that are relevant for the city landmarks and negative examples from totally unrelated Flickr groups. The representative tags are extracted and used to construct a photo summary. We evaluated our algorithms with the help of a user study. Its results show that this new method outperforms state-of-the-art geo-tagging based systems on a large number of cities. Detailed evaluation and explanation of this work can be found in (Abbasi et al., 2008)

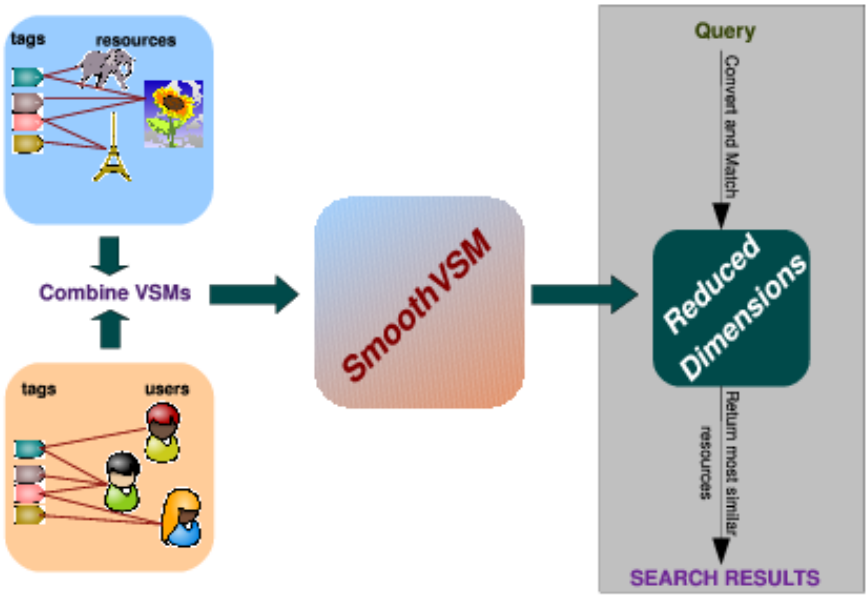

Figure 3.1: Overall process of *Triple Play*.

To evaluate the classification results by inferring knowledge from Flickr groups, we plan to recruit 20 volunteers who will be expert users and familiar with photo sharing and search services. The pool of queries will consist of 50 large cities around the world. Each user will be asked to evaluate two result sets for 10 randomly selected cities, where each city will be picked once for 4 users. Two summaries will be mixed on a single screen, with one result set created using our algorithm and one coming from World Explorer API (which identify those landmark photos which have geographical co-ordinates associated to them). The users will not know which photo is coming from where and result sets from two systems will be randomly interleaved. As a city overview, each time 60 photos with photo title and first 5 tags will be displayed. Users will be able to decide whether the displayed photo is a “sight", “non-sight", and or they “don’t know" about it. User would also be asked about the overall results for each method, just to confirm that users’ judgments on separate photos are consistent with users’ overall impression. The experiment will take about 30 minutes per user.

3.3 Sound Classification

Sony CSL conducted intensive and rigorous experiments on the automatic classification of acoustic signals with respect to their associated tags. This is a supervised-learning task that is typically addressed by training a classifier for each tag. The classifiers are trained on feature values computed for each title and they learn the tags given by humans (the so-called ground-truth). The performance of these *acoustic* classifiers are rarely satisfactory.

We therefore have introduced the correction hypothesis, which postulates that it is possible to exploit the redundancies between tags to correct some of the errors of acoustic classifiers. In our correction approach, we train a *correction* classifier for each tag using the output of all the acoustic classifiers.

We conducted a series of experiments to validate this hypothesis on a large-scale database of music and metadata (32,000 titles and 600 attributes per title). In order to avoid biases due to the choice of the feature set, we have trained and tested several classifiers using two distinct feature sets. The first set, called *generic*, comprises audio features from MPEG7-audio. The second set, called *specific*, consists of Sony’s proprietary features, designed for high-level music categorization

and used in consumer electronic products (e.g. hard-disk based HiFi systems).
We drew the following conclusions from these experiments (Pachet and Roy, 2008a,b,c; Rabbat and Pachet, 2008):

- **Acoustic classifiers perform much better than random oracles:** We observed that for most of the tags, the corresponding individual acoustic classifier performs substantially better than a random oracle would
- **The performance of acoustic classifiers follows a power-law:** The distribution of the performances (in *min-F-measure*) of the acoustic classifiers vary from 0% for both feature sets to 74% for the generic features and 76% for the specific ones. The log-log graph in Figure 3.3 shows that the statistical distribution of the performances is close to a power law distribution.
- **The specific features perform slightly better than generic features.**
- **The correction hypothesis is true:** On average, correction classifiers perform better than the corresponding acoustic classifiers (Fig. 3.3).
- **The improvements are feature set independent:** An interesting result of this study is the parallelism between the performance improvements with the two feature sets.
- **Semantic categories are not grounded:** There are no relation between the semantic category of the tags and their ”groundednesss”. Each category contains both very good and very bad tags. For instance, the “Style” category contains the tags ”Urban” and ”Rock”, which are in the top list of the acoustic classifiers but also ”Folk Rock” which yields extremely bad results.

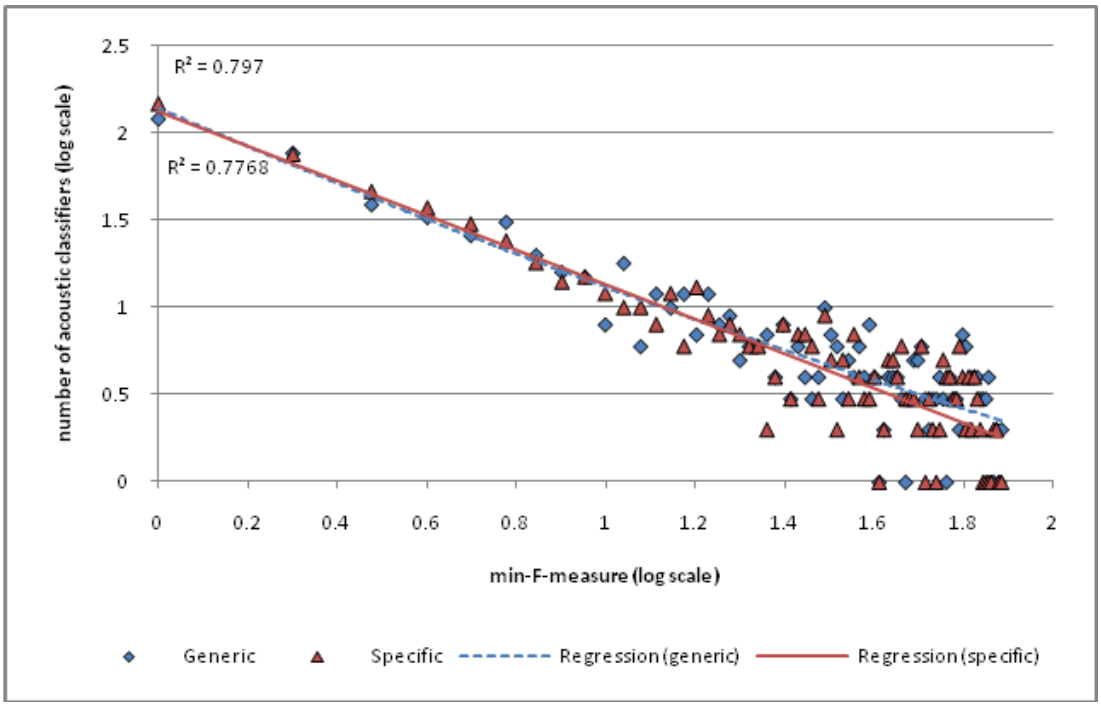

Figure 3.2: Log-log graph of the distribution of the performance of acoustic classifiers for both feature sets.

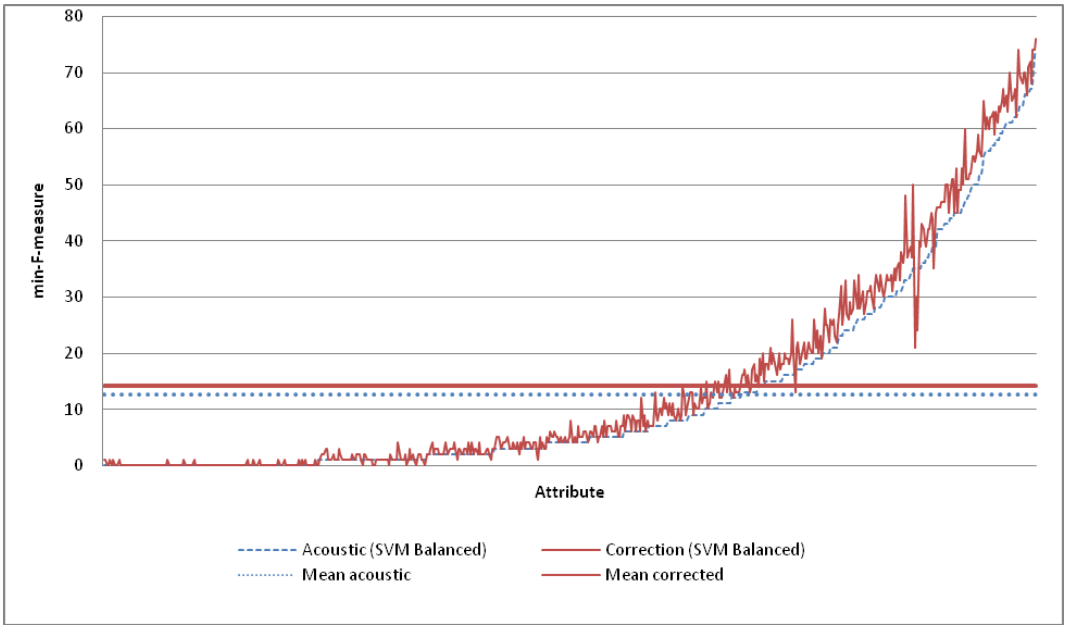

Figure 3.3: The improvement in classifier performance (specific feature set). x-axis: attribute indices, sorted by increasing performance of the corresponding acoustic classifier. y-axis: the min-F-measure of the classifiers. The performance of the acoustic classifiers is the "smooth" dashed line. The two horizontal lines represent the average performance of the acoustic classifiers (dashed line) and of the correction classifier (plain line).

3.4 Semantic Grounding

Formal representations of knowledge were leveraged to gain insights into the semantic aspects of folksonomy-based notions of tag and resource similarity. Measures of tag similarity (Cattuto et al., 2008a,b) were grounded in Wordnet (`http://wordnet.princeton.edu/`), and measures of resource similarity (Markines et al., 2008) were grounded in the Open Directory Project (DMOZ). Semantic grounding allowed us to provide a semantic characterization of statistical measures on the folksonomy.

As an example, consider a distributional notion of tag similarity based on tag co-occurrence. In this case, two tags are considered "similar" is they co-occur in similar ways with all the other tags in the system, i.e., if they have similar co-occurrence "fingerprints" or "contexts". Fig. 3.5 displays the co-occurrence profiles for a few tags, and shows that synonym tags that differ only in the morphology (*games* and game) have strongly similar co-occurrence profiles. This notion of distributional similarity can be characterized more precisely by mapping pairs of similar tags to Wordnet synonyms and looking up, in Wordnet, the semantic relation between the pair of mapped tags. The result of this procedure is shown in Figs. 3.6 and 3.7. One can clearly see that tag-based distributional similarity (marked as *TagCont* in the figures) yields a large fraction of actual synonyms (as recognized by Wordnet) and a comparatively high fraction of tags that are two edges away from each other in the Wordnet subsumption hierarchy. Analyzing the composition of these paths (Fig. 3.7) shows that they are all made up by one edge towards a hypernym and one edge towards a hyponym, i.e., they lead to sibling tags in the Wordnet hierarchy.

In general, several kind of similarity notions can be defined, based on different projection and aggregation schemes. We systematically characterized entire classes of similarity measures by using procedures of semantic grounding like the one outlined in the above example. Pair-wise similarities given by different measures were compared by using Kendall's tau measure of correlation, as shown in Figs. 3.8 and 3.9. Our analysis showed that measures of similarity based on mutual

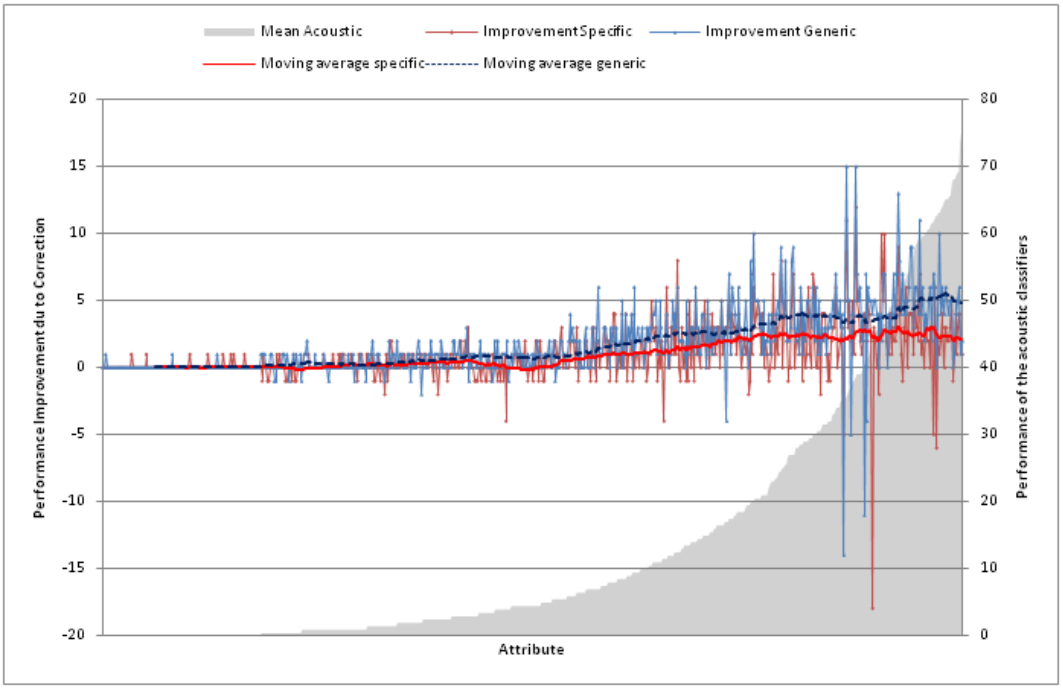

Figure 3.4: The performance gain of the correction classifiers as a function of the performance of the associated acoustic classifier. The red (resp. blue) curve is the increase in performance of the correction classifier over the acoustic classifier when using specific (resp. generic) features, the scale is reported on the left vertical axis, in percentage increase of the min-F-measure. The brown bars show the performance (averaged on both feature sets) of the original acoustic classifier. The scale is reported on the right vertical axis, in min-F-measure. Those curves were computed on overlapping series of 50 attributes of increasing performance.

information (of resource/tag vectors) afford the highest accuracy, but this comes with the price of high computational complexity, and makes those measures hard to deply in scenarios involving large-scale folksonomis. However, a scalable strategy can be devised: expensive measures of similarity can be computed on a per-user basis, and then aggregated over users with a simple form of collaborative filtering. As shown in Fig. 3.10, this strategy makes the incremental update of the similarities scalable, as it doesn't require recomputing all the similarities when new triples flow into the system.

Summarizing, our research led to two main results, presented in Ref. (Markines et al., 2008):

- a clear definition of classes of folksonomy-based measures that involve different levels of projection and aggregation, symmetrically defined on both tags and resources (and in principle, users);
- an empirical analysis of the computational scalability of the measures under investigation, showing that by suitably choosing projection and aggregation schemes it is possible to devise measures that are both semantically accurate *and* scalable.

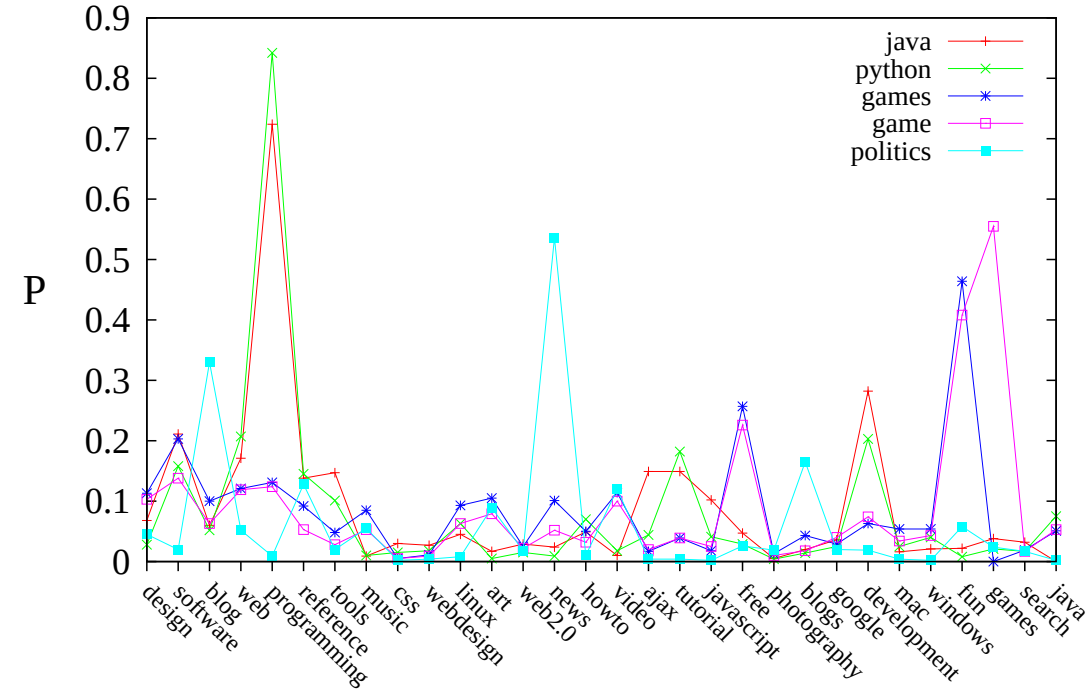

Figure 3.5: Tag co-occurrence fingerprint of five selected tags in the first 30 dimensions of the tag vector space.

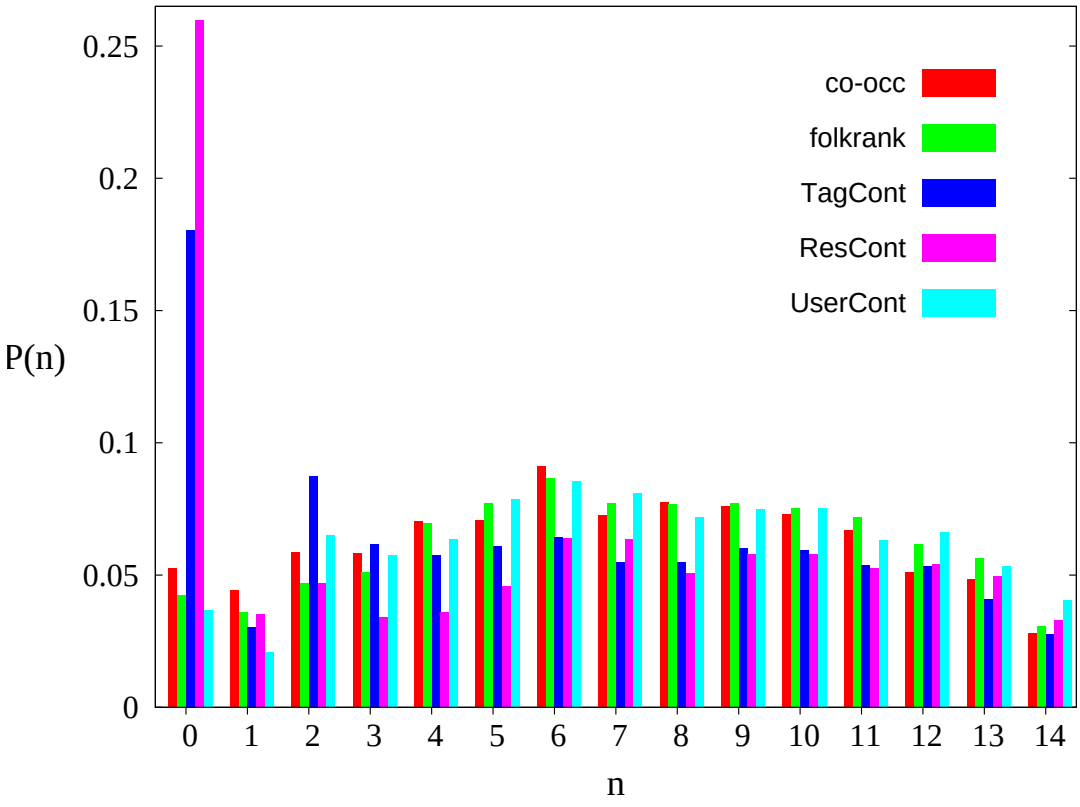

Figure 3.6: Probability distribution of the lengths of the shortest path leading from the original tag to the most closely related one, as defined according to several different notions of similarity. Path lengths were computed using the subsumption hierarchy in WordNet.

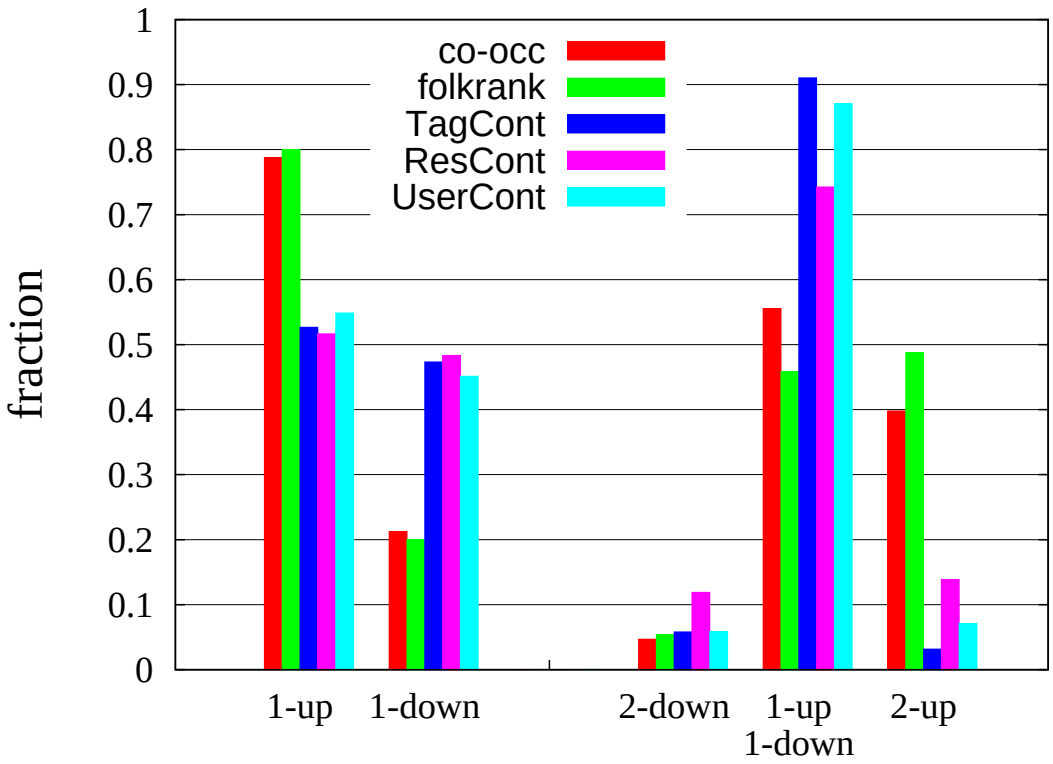

Figure 3.7: Edge composition of the shortest paths of length 1 (left) and 2 (right). An “up” edge leads to a hypernym, while a “down” edge leads to a hyponym.

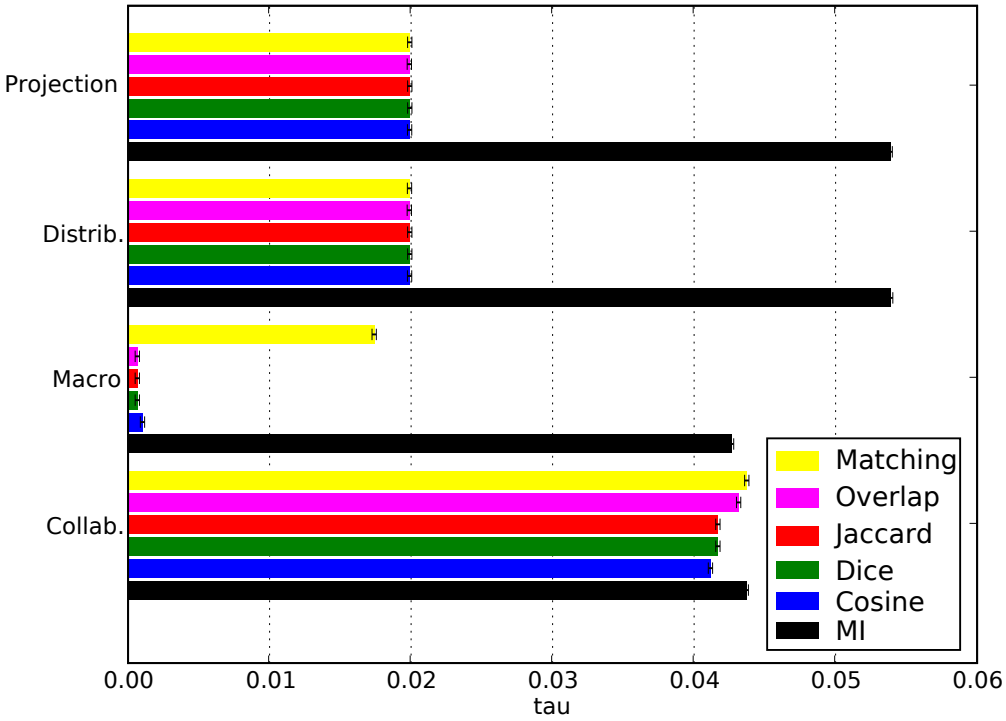

Figure 3.8: Tag-tag similarity accuracy, according to Kendall's τ correlations between the similarity vectors generated by the various measures and the reference similarity vector provided by the WordNet grounding measure (Jiang-Conrath).

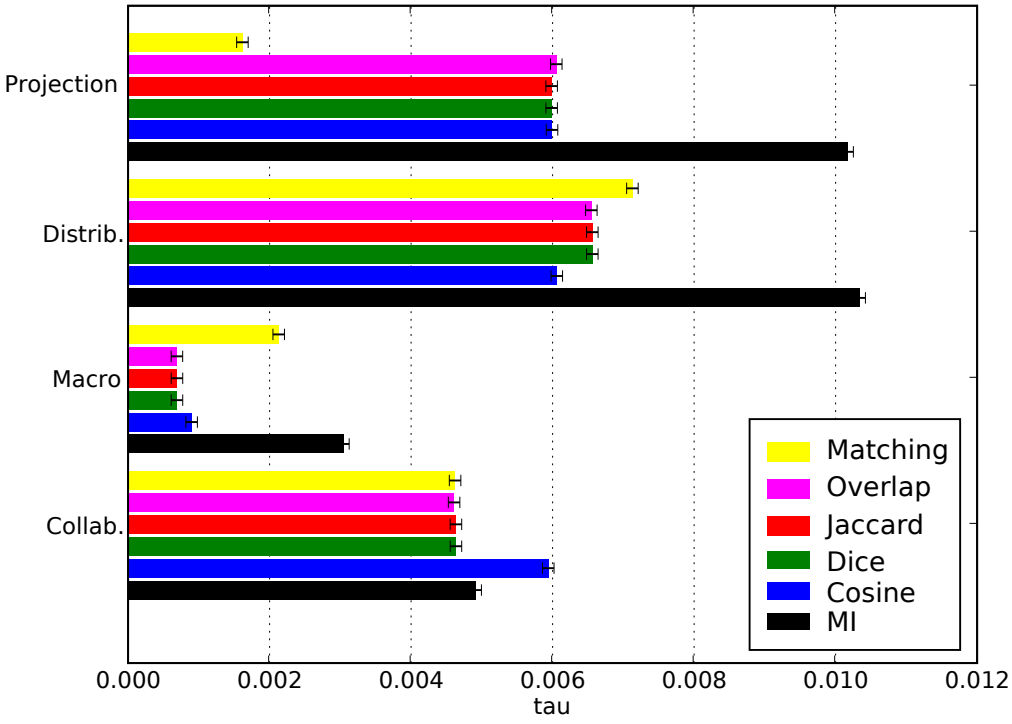

Figure 3.9: Resource-resource similarity accuracy, according to Kendall's τ correlations between the similarity vectors generated by the various measures and the reference similarity vector provided by the ODP (DMOZ) grounding measure (a modified form of Lin's measure).

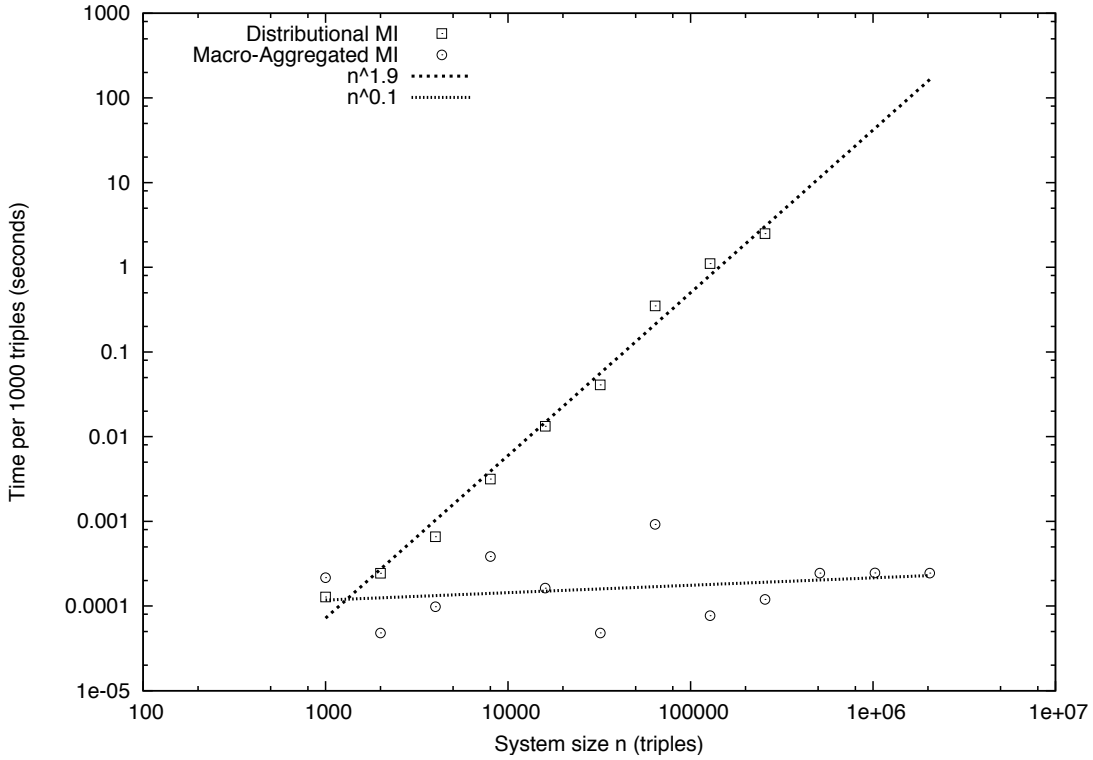

Figure 3.10: Scalability of the mutual information computation of resource similarity, for different aggregation methods. We display the CPU time necessary to update the similarities after a constant number of new annotations are received, as a function of system size n. Best power-law fits $time \sim n^{\alpha}$ are also shown.

Chapter 4

Conclusion

In this deliverable we have presented our results concerning the use of background knowledge for semantic inferencing. We identified different sources of background knowledge and two main application cases. The first one is the pre-processing of the datasets by removing noisy data or enhancing the data with additional knowledge such as synonyms, word translations etc. A concrete filtering architecture was described and we observed that useful sources for background knowledge are, for example, Wordnet, Wikipedia and Google search.

In the second part dfferent approaches for semantic inferencing were shown, not all of them actually using background knowledge. The Triple Play approach aims at finding hidden concepts by applying Latent Semantic Analysis on “smoothed” vector space models which are derived from recombining the original item relations. Furthermore, two methods for classifying images and audio data were explained. The first one identifies location information used in image tags while the latter one organizes audio data in classes by their characteristics and explicit annotations.

The experiments and evaluations conducted with the algorithms clearly show the benefits of applying background knowledge since it is clearly possible to improve the results in comparison to the solutions were no background knowledge was applied.

Bibliography

Rabeeh Abbasi and Steffen Staab. Introducing triple play for improved resource retrieval in collaborative tagging systems. In *In: Proc. of ECIR'08 Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2008)*, 3 2008. URL `http://www.uni-koblenz.de/~abbasi/publications/Abbasi2008ITP.pdf`.

Rabeeh Abbasi, Sergey Chernov, Wolfgang Nejdl, Raluca Paiu, and Steffen Staab. Exploiting Flickr Social Information for Finding Landmark Photos. *Submitted in Proc. CIKM*, 2008.

Ivan Cantador, Martin Szomszor, Harith Alani, Miriam Fernández, and Pablo Castells. Enriching ontological user profiles with tagging history for multi-domain recommendations. In *Proc. Workshop on Collective Semantics: Collective Intelligence and the Semantic Web (CISWeb 2008), in 5th ESWC, Tenerife, Spain*, 2008.

Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic analysis of tag similarity measures in collaborative tagging systems. In *Proceedings of the 3rd Workshop on Ontology Learning and Population (ECAI2008)*, 7 2008a. URL `http://arxiv.org/abs/0805.2045`. http://arxiv.org/abs/0805.2045.

Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic grounding of tag relatedness in social bookmarking systems, 2008b.

Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Gerd Stumme. Social similarity, 2008.

F. Pachet and P. Roy. Analytical features: a knowledge-based approach to audio feature generation, 2008a. Submitted to Journal of Artificial Intelligence Research.

F. Pachet and P. Roy. Is hit song science a science?, 2008b. Accepted to the International Symposium on Music Information Retrieval (ISMIR).

F. Pachet and P. Roy. Improving multi-class analysis of music titles: A large-scale study, 2008c. Accepted with major changes, to appear in IEEE Transactions on Audio, Speech and Language Processing.

P. Rabbat and F. Pachet. Direct and inverse inference in music databases: How to make a song funk?, 2008. Accepted to the International Symposium on Music Information Retrieval (ISMIR).

Martin Szomszor, Harith Alani, Ivan Cantador, Kieron O'Hara, and Nigel Shadbolt. Semantic modelling of user interests based on cross-folksonomy analysis. In *submitted to Int. Semantic Web Conf., Karlsruhe, Germany*, 2008a.

Martin Szomszor, Ivan Cantador, and Harith Alani. Correlating user profiles from multiple folksonomies. In *Proc. Int. Conf. Hypertext (HT08), Pittsburgh, PA, USA*, 2008b.