

Collaborative Lightweight Ontologies in Open Innovation-Networks

Walter Christian Kammergruber¹, Michele Brocco¹, Georg Groh¹ and Manfred Langen²

¹ Technische Universität München, Boltzmannstr. 3
85748 Garching, Germany

WKammergr@gmail.com, brocco · grohg@in.tum.de

² Siemens AG, Corporate Technology

Otto-Hahn-Ring 6

81739 Munich, Germany

Manfred.Langens@siemens.com

Abstract. Social tagging is a simple, flexible and widely used method for annotating almost every kind of entity. In the context of open innovation tagging can be used to manage and structure competencies and topics. We describe an approach where tagging data can be enriched with lightweight structures to result in a thesaurus. We describe three different methods that lead to these lightweight structures: Statistical analysis of social tagging data, mapping tags to existing ontologies or other structured input and collaborative tag thesaurus creation. The resulting structures can be used in a further step for query refinement in the case of tag-based search. In addition they can serve as base for tree-like navigation interfaces.

1 Introduction

1.1 Open Innovation

“Open innovation is the use of purposive inflows and outflows of knowledge to accelerate internal innovation, and expand the markets for external use of innovation, respectively. [This paradigm] assumes that firms can and should use external ideas as well as internal ideas, and internal and external paths to market, as they look to advance their technology.” [4]

Through the openness of this process ideas or technologies from outside a firm can be brought into product development whenever competencies or resources are not available inside the firm’s boundary. Competencies play a central role in open innovation processes: Especially radical innovation may imply a new type or mix of competencies. Hence, the set and structure of competencies in a network of innovators may rapidly change over time. Beside heterogeneous competences, a heterogeneous set and structure of ideas has to be managed in open innovation. Idea contest or sometimes called “innovation jams” for example generate a huge amount of input that has to be dealt with in a proper way.

Formal approaches for structuring competencies or ideas such as ontologies are complex to design and maintain. Therefore managing ideas and competencies (and especially managing their structure and relations) in open innovation is a major challenge which can be faced by developing appropriate supporting mechanisms.

1.2 Social Tagging

Social Tagging has become a very popular tool for categorizing knowledge items over the last few years [21,22]. One of the first social tagging applications was Delicious, a social bookmarking platform. Many other Internet services such as Flickr, Last.fm, 43things, Citeulike or Bibsonomy³, make use of tagging for structuring their content.

In most current work social tagging data is also referred to as folksonomy [32]. A folksonomy is a portmanteau consisting of the words taxonomy and the word folk. The term refers to the implicit structures (therefore taxonomy⁴) between entities emerging from the individual tagging behaviour (therefore folk).

We define a folksonomy similar to the definition given by Hotho et al. [15]:

Definition 1 (Folksonomy). *A folksonomy is a tuple $\mathbb{F} := (U, T, R, Y)$ where*

- *U , T and R and R are finite sets, whose elements are called users, tags and resources, resp.*
- *Y is a ternary relation between them, i. e. $Y \subseteq U \times T \times R$, called tag assignments.*

In general a folksonomy contains multinary relations between user, tags and resources. Additionally sometimes the system, meaning the tagging application in which the individual tag assignment occurred, is also considered [10]. For the sake of simplicity we intentionally exclude the system part in this paper.

1.3 Collaborative Construction of Lightweight Ontologies in OI

Open Innovation, as described by Chesbrough's paradigm, may involve a whole network of innovators inside and outside a company. Thus, it can be said, that an open innovation process is a sort of social process. Contributions for the development of an innovation may come (collaboratively) from communities, individuals and/or organization, all together.

For the collaboration adequate collaborations and communication tools have to be provided. Additionally supportive OI-tailored tools can help to foster and guide the innovation process. However, often these tools need complex problems to be solved, such as the evaluation of contributions in order to recommend

³ <http://www.flickr.com/>, <http://www.last.fm/>,
<http://www.43things.com/>, <http://www.bibsonomy.org/>

⁴ Taxonomy actually can be considered as the wrong expression since there are no real hierarchical structures present.

experts in the OI network. Often no algorithms or automated approaches for such problems can be found or formulated. In that cases, these problems can be solved by (parts of) the OI community itself and by wisdom of the crowd.

One of these tasks is the aforementioned (section 1.1) management and structuring of ideas and competencies that we address by the paradigm of collaborative lightweight structures. Similarly, we are looking for an approach for structuring other kind of items and entities (e.g. topics).

Our target community for our investigation is the KoPIWA Open Innovation community, which consists of small and medium enterprises of the digital economy which are members of the organization BVDW in Germany. The original Open Innovation Platform is a Liferay⁵ based community portal, with standard collaboration tools included (wiki, blogs, forums, etc.).

In order to implement our ideas we extended the platform with an additional tagging system which comprises the tagging of competencies and topics for each contribution respectively. Contributions, in our case, include posts in forums, blogs and wikis as well as uploaded files, documents or announcements. Based on this tagging system we want to develop a concept to structure competencies (and topics) in a lightweight structure.

In the next sections we will first introduce social tagging in a more general way and present different types of structures for tags. After that, we will briefly discuss existing automatic approaches for structuring tags and motivate why they are not satisfactory. Based on this analysis, we will present an approach which mixes automatic methods for structuring tags with human interaction.

2 Problems with Tagging

The main advantage of (social) tagging over traditional annotation methods (e.g. modeling and instantiating entities in ontology- or taxonomy-based systems) is the ease of use. Not only experts but also untrained users can utilize tagging for their needs. In tagging there is no restriction concerning the allowed terminology. Social tagging, as concept within Web 2.0, supports the interaction of users on the social web because tags are not only intended for personal use, but are also intended for others to give them the opportunity to quickly estimate semantic aspects of given information items.

Because of its simplicity tagging in its basic form lacks any form of explicit structure that comes with other more formal categorization methods (e.g. ontologies or thesauri).

In general, one can distinguish between two categories of problems. The first category contains very common problems that come with free text annotations in general.

- *Typos, spelling mistakes or different spellings*: This is the simplest case where tags are susceptible. A user might type “instuments” instead of “instruments”.

⁵ <http://www.liferay.com>

Also spelling variants, e.g. between American/ British English such as “color” and “colour” are a problem.

- *Special chars for word combinations* (“_”, “.”, “/”) or *camelCase*: Depending on the individual taste of a user or sometimes related to restrictions given by the tagging application (delicious does not allow white spaces in between tags; atlassian confluence⁶) “open source”, “open_source”, “openSource” might be a user’s choice to tag a specific item associated with open source software.
- *Meta-Noise*: In Internet tagging applications there also effects related to SPAM (e.g. in delicious user accounts are abused to link to dubious sites in order to retrieve more attention), trolls (e.g. some user try to miscategorize items for fun or destructive reasons) and pseudo ‘experts’ (some people overestimate the expert knowledge)
- *Different languages*: In tagging application with international users variants of the same term may occur in different languages, e.g. one might find pictures somehow related to “luck” under “Glück” (German) as well as “suerte” (Spanish) or “bonheur”(French).

These problems are also targeted in classical information retrieval [19]. Syntactic issues can be dealt with by using spell checkers, stemming algorithms (e.g. Porter stemming [26]) or comparing string distance metrics (e.g. Levenshtein distance [2]). SPAM detection in folksonomies for example is discussed in [17]. Language detection might be done e.g. by matching tags against several dictionaries.

Our second identified category of problems is the lack of structure:

- *Synonyms*: Two words are synonym when they have the same (or nearly the same) meaning. Examples are “buy” and “purchase” which can be nearly interchangeable used. “dog” and “Canis familiaris” are synonym as well, but are normally used in a different context – the first one as widely used term, the second one mostly used in scientific articles.
- *Homonyms/ Polysemy*: Homonym means that two words are spelled (homograph) or pronounced (homophone) in the same way [7]. For our case we are only interested in homographs. A typical example is “bow” can have different meaning such as the weapon or to bend forward.
- *Acronyms*: Acronyms are abbreviations of typically longer terms. “GIS” can stand e.g. for “Geographic information system”, “Greenland ice sheet” or “Gruppo di Intervento Speciale”
- *Level of abstraction – hyponyms or hypernyms*: Depending on the expertise of an user or other circumstances (e.g. who is tagging for who) different levels or abstractions for the chosen tags can be used. A picture of an angora cat can be tagged e.g. with “angora cat”, “cat”, “mammal”, “animal” or “creature”.

The second category of problems include issues that are related to missing semantic structures in social tagging data. Folksonomies can be viewed as a soft collaborative classification or structuring scheme based on a common semantic

⁶ A wiki software <http://www.atlassian.com/software/confluence/>

2. *Navigation:* In general folksonomies with many tags can not be easily explored by browsing through some kind of network or tree structure as it is the case with taxonomies.

When tags are considered as a special form of manual indexing [31] these two use cases are the most straightforward ones. A fundamental question is therefore how social tagging data can be enhanced with semantic relations in order to achieve navigation support and enable semantic search.

3 From Folksonomies to Thesauri

A thesaurus is a controlled vocabulary of terms that can be used as keywords. There are several variants of thesauri depending on the area they are used in. Peter Mark Roget's famous *Thesaurus of English Words and Phrases* (1852) initiated the concept of a linguistic thesaurus. A linguistic thesaurus is some kind of dictionary where words are arranged systematically. This type of thesaurus is the most widely used and known one since it is included in popular word processors such as Microsoft Word or Open Office. The use case of a linguistic thesaurus is hereby assisting user in finding alternative words for avoiding repetitions of phrases when writing texts. Alternatively a linguistic thesaurus can be used to determine the meaning of a term when in doubt.

In information science or in the context of libraries a thesaurus is used to categorize information objects. A thesaurus is some kind of classification system backed by a controlled vocabulary with several kinds of relation between terms contained in the vocabulary. Here a thesaurus helps conducting research to a certain topic, e.g. by allowing a user to enter a classification system through different terms having the same meaning.

Thesauri are furthermore used in several scientific fields such as biology or medicine. These thesauri typically have a very narrow domain and a well defined and restrictive terminology. Sometimes these thesauri are a preliminary stage to an ontology and also referred to as one. The Radlex ontology⁸ is an example for such an ontology/ taxonomy/ thesaurus. It is a controlled vocabulary to classify information items, such as x-ray images or medical reports, in the area of radiology.

In context of information retrieval typically a thesaurus is used to alleviate problems resulting from trivial search variation or from term ambiguity by offering terminology control [30].

As described in section 2, folksonomies lack explicit formal structures. Therefore our goal is to extend a folksonomy with relations and generate a thesaurus based on tags.

Thesauri allow suggesting alternative search terms/tags or can be used for automatic query expansion. An advantage of having folksonomies enriched with thesaurus relations is the ability to explore the explicit tree/network structure

⁸ <http://www.rsna.org/radlex/>

to find resources. This is especially useful if one does not know with what exact tag the sought item has been labeled. Additionally serendipity effects may arise.

Some popular thesaurus relations are defined in the thesaurus standard ISO 2788. We believe for our case the most useful relations are *Use synonym*, *Broader term*, *Narrower term*, *Related term*. These relations can be employed in query expansion or refinement depending on the size and the quality of the result list. Additionally we use a relation *Ignore Relation* that enables the user explicitly state that there is no relation between two tags. This is useful when automatic algorithm produce false or undesired relations. Having too many types of relations may confuse untrained user and may lead to the same problems that come with classical ontology engineering processes.

In the following sections we will first introduce methods for extracting structural information out of folksonomy through statistical analysis methods. Then we will sketch approaches for mapping tags to an existing ontology. In the last section we describe a thesaurus editor where user can define relations between tags manually.

3.1 Statistical Analysis

One very popular approach for finding relations between tags is using the co-occurrence of tags, meaning two tags have been used together to annotate an object. This method is rather simple and can only deliver some kind of unspecific relation. The exact kind of semantic relation between two tags is very hard to determine and depends on the actual tagging practices of the single user in a tagging application.

We start by defining a utility function called *cover*. The function collects all user-resource tuples from all tag assignments where at least one user has applied a tag to a single resource. The frequency of a tag A in a tagging system is equivalent to the cardinality of $cover(A)$.

Definition 2 (Cover). *Let $A \in T$ be a tag, then*

$$cover(A) = \{(u, r) \in U \times R \mid \exists u \in U : (u, t, r) \in Y \wedge t = A\}$$

defines the finite set of user-resource tuples that have been tagged with A .

Having *cover* defined the absolute co-occurrence of two tags, meaning two tags have been used together in a tag assignment can be defined as followed:

Definition 3 (Absolute Co-Occurrence). *Let $A, B \in T$ be tags, then the absolute co-occurrence AC is defined as:*

$$AC(A, B) = |cover(A) \cap cover(B)|$$

This is the most popular approach for computing relations between tags in recent work – probably because it is easy and efficient to compute. Its major drawback is the fact that if A and B are very frequent tags in typically get higher common

frequency values than they would get if they were rarely used. This might lead to distorted results in the interpretation of the strength of a co-occurrence relation.

An alternative method for computing relations between tags can be formulated with the relative co-occurrence. In the relative co-occurrence the frequency of the individual tag is also taken into account. The relative co-occurrence is a special form of the Jaccard similarity coefficient [11].

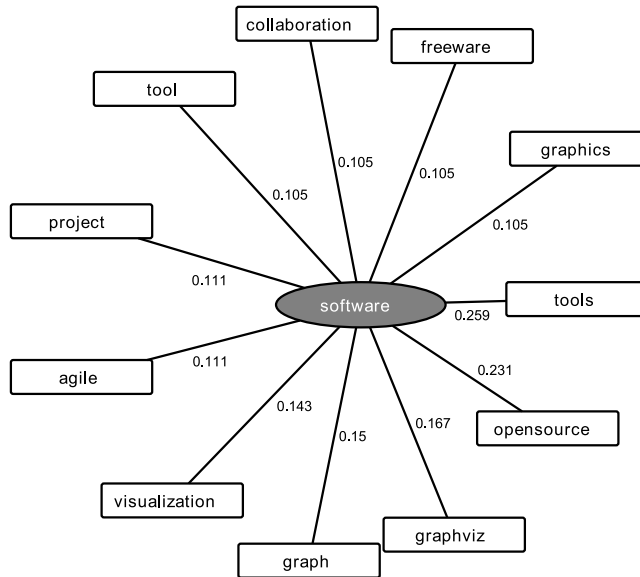


Fig. 2. Tag Relations: Eleven tags with the highest RC value to the tag “software”. They are ordered clockwise beginning with “tools”. The length of the edges a proportional to the similarity of to tags.

Definition 4 (Relative Co-Occurrence). Let $A, B \in T$, then the relative co-occurrence RC is defined as:

$$\begin{aligned}
 RC(A, B) &= \frac{|cover(A) \cap cover(B)|}{|cover(A) \cup cover(B)|} \\
 &= \frac{|cover(A) \cap cover(B)|}{|cover(A)| + |cover(B)| - |cover(A) \cap cover(B)|}
 \end{aligned}$$

A potential semantic similarity between two tags A and B can be estimated by the corresponding $AC(A, B)$ or $RC(A, C)$ value — the higher the RC value of two tags the stronger the potential semantic proximity of those two tags.

Fig. 2 is an example for similar tags based on the relative co-occurrence. The graph is deduced from the tagging practise of a delicious user. In the center is

the tag “software” and in clockwise order (beginning with “tools”) the eleven most similar tags determined by the RC value are displayed.

Grahl et al. [9] use a standard clustering algorithm (KMeans - see [6]) to find hierarchies in sets of tags. They use an iterating approach where first a coarse clustering with 300 clusters is computed. In a second step each resulting cluster is clustered into 20 clusters. In a last step there are the two most representing tags from the clustering in the second step merged into another clustering. Other examples for clustering tag sets can be found in: [29,28,13].

Cattuto et al. investigate several methods for automatically discovering relations between tags in social tagging applications [3]. They test three different approaches: tag co-occurrence, cosine similarity of co-occurrence distributions, and FolkRank [14]. In order to provide a semantic grounding of their folksonomy based measures, they are trying to map tags to synsets of WordNet. They compare the semantic similarity computed from WordNet with the ones determined through folksonomy based similarity measures.

A newer work by Markines et al. evaluates different similarity measures for emergent semantics of social tagging data [20]. First they introduce a mapping/projection of the multinary relations represented by a folksonomy to a simple matrix. Based on the different resulting matrices they evaluate several similarity measures (such as Jaccard or Cosine) against WordNet and DMOZ.

3.2 Mapping Ontologies/ other Structured Input to Tags

A lot of methods are conceivable for mapping tags to concepts in an ontology. In general a possible solution uses a combination of string distances, stemming algorithms and comparison of graph structures – both of the ontology on the one hand and the relations between users, tags and resources in a folksonomy on the other hand. As previously mentioned (section 3.1) Cattuto et al. compare automatically derived semantic similarities between tags of a folksonomy with the similarity of the corresponding concepts computed from the given graph structure in WordNet. Not mentioned in this paper is, how the actual mapping between tags and concepts was achieved. They only state that “roughly 61% of the 10 000 most frequent tags in del.icio.us can be found in WordNet”.

Al-Khalifa et al. follow an approach where tags are first normalized, i.e. stemming algorithms are applied, then tags are grouped and general tags are removed [1]. In a last step the resulting stemmed tags are mapped to (stemmed) concepts of an existing ontologies.

Laniado et al. [18] investigate how an ontology (for their case the *noun hierarchy* of WordNet) can be integrated into a navigation interface for an existing folksonomy. When it comes to mapping tags to concepts in WordNet they state that only 8% of the different tags in their data sample (480,000 different tags collected from 30,000 del.icio.us users) find a corresponding concept in WordNet. Regarding the most popular tags they observe a higher percentage of matches. For tag disambiguation (homonyms) they use a semantic similarity metric based on the work of Pedersen et al. [24].

Depending on the kind of desired structure information – for our case the described thesauri relations – not only full-blown ontologies but also thesauri such as WordNet (Princeton University – English), Wortschatz (University of Leipzig – German) or GermaNet (University of Tübingen – German) and other structured input such as DMOZ⁹/ Google Directory¹⁰ or Dbpedia¹¹ can contain valuable structure information. Though structured sources may be a valuable input, the mapping between tags and items from the sources are most likely incomplete and error-prone to some point.

3.3 Thesaurus Editor

We believe that the results of statistical analysis of folksonomies or the mapping of tags to ontologies are not accurate enough to generate a suitable thesaurus. Automatically computed similarities may contain errors depending on the folksonomy data, e.g. as a result of different tag usage patterns [8]. Mapping tags to concepts of an ontology is equally error-prone since normally not all tags are contained as concepts in an ontology. Additionally ambiguities of terms (e.g. homonyms or acronyms) might not be resolvable. In contrast a in general more precise manual creation of thesauri (following a formal procedure) is expensive and time consuming. Therefore our proposed approach brings together the two methods into a combined, semi-automatic approach.

Summed up there are three categories of relations between tags:

- statistically computed ones,
- relations found through mapping tags to concepts in an ontology, and
- the manually defined relations between tags.

To determine a semantic similarity between tags each of these kind of relation can be differently treated and taken into account. We consider the manually defined relations between tags as the most valuable ones since a user has defined them. Manually created relations are less likely to be wrong than algorithmically inferred ones.

The results of the statistical analysis and ontology mapping are only suggestions for relations between tags. A user decides whether a proposed relation is correct or not. Only verified relations are included in the final thesaurus.

To make this process as easy as possible, a user can formulate thesaurus relations through web based thesaurus editor with a drag and drop style interface. With this editor she can express her personal opinion that one tag has some certain relation to another tag, e.g. “ajax” can be a narrower term of “web2.0”. By proposing tag relation based on the described automatic methods the process of creating a thesaurus is simplified since in many cases the user only has to confirm tag relations and does not have to think about these relations on her own – though she is able to express additional relations between tags.

⁹ <http://www.dmoz.org/>

¹⁰ <http://www.google.com/dirhp>

¹¹ <http://www.dbpedia.org>

The thesaurus editor enables the user to extend her own personal tag space with more structure. These relations can be a very personal view with which another user might disagree. But in contrast, given many user have formulated the same relation between two tags, we assume with some certainty that this relations may be correct.

As already mentioned we assume that the most useful relations are:

- *Use synonym*: The tags can be used interchangeable. In general one can distinguish different levels of synonymy. Words can have the exact same meaning or only in some context. We do not separate those cases for sake of simplicity. An example for synonyms is “person” and “individual”.
- *Broader term*: A tag has a more general meaning than an other term. E.g. “mammal” is a broader term of “primate”.
- *Narrower term*: A tag has a narrower meaning than an other term. E.g. “primate” is a broader term of “mammal”.
- *Related term*: This is the weakest relation. Two tags are only related in some way, e.g. “web2.0” and “ajax”.

Additionally we define an artificial relation with the name *Ignore Relation*. This relation allows the user explicitly express that there is no relation between two tags. Automatic algorithm may infer undesired relations that a user can then dismiss.

Fig. 3 shows a screenshot of the web interface. A user can define via drag and drop relations between tags. In the example the tag “knowledgemanagement” is selected (2). Selecting a tag can be done by double clicking a tag in any tag boxes. There is a simple filter mechanism for searching for tags (1). A users starts typing and the resulting tags are displayed according to entered letters. The resulting tags can be set as current tag or dragged into the relation boxes (3). The relation boxes are used as drop zones. Where a user can define a relation between the current tag and another tag. She simply drags the tag in to the desired box. The type of boxes match our chosen thesaurus relations enumerated above. If the user has already defined relations between the current tag and other tags they are filled in to the relation boxes accordingly. Utilizing the two algorithmic approaches to find relations between tags we display computed relations between the current tag and other tags in several suggestions boxes (4). In the first box we display tags that have a low string distance to the current tag. For the example knowledgemanagement the results are not optimal but the string distance can be used to find synonyms with spelling variants or singular/ plural. “event” and “event” or “web_2.0” and “web2.0” can be listed as examples. Another box contains the co-occurrence matches (in the example “km”, “wissensmanagement” and so on). In the most left box suggestions generated by mapping a tag to terms in external structured input are displayed.

Fig. 4 describes the underlying data model. We follow the popular n-Array Relations Design Pattern¹². Sometimes this kind of artificial classes are called

¹² <http://www.w3.org/TR/swbp-n-aryRelations/>

You are here: > Home > Thesaurus Editor

[1]
filter:

productivity
pilot
participationage
plm
podcast
processes
ptech
Personal and Distributed Knowledge Management
personal
pictures
platform
pm
portal
presentation
prjalexandria
problems
productivitykillers
project management
psychology
ptools

Current Tag: **knowledgemanagement** **[2]**

Relations: **[3]**

Synonym: km, wissensmanagement
Related: blogging, enterprise2.0, knowledgework, weblogs
Broader:
Narrower:
Ignore:

Suggestions: **[4]**

Spelling Variant: cortenemanagement, project management
Related Terms: km, wissensmanagement, web2.0, enterprise2.0, expertnetworks, infoman, knowledgework, benchmarking, blogcases, call, complexity, distributed, history, infom, knowledge, ktypes, kybernetik, presentation
Other Relations: Knowledge value, Personal knowledge management, Project blog, Collective intelligence

Fig. 3. Tag Thesaurus Editor example: A user can define relations via drag and drop to a relation box (synonym, narrower, broader, related (3)) between a selected tag (in this example knowledgemanagement (2)) and another tag. Her already defined tag relations are displayed in the boxes. For the current tag she gets suggestions for possible related tags (4). Additionally a user can search the folksonomy by applying a simple filter mechanism (1).

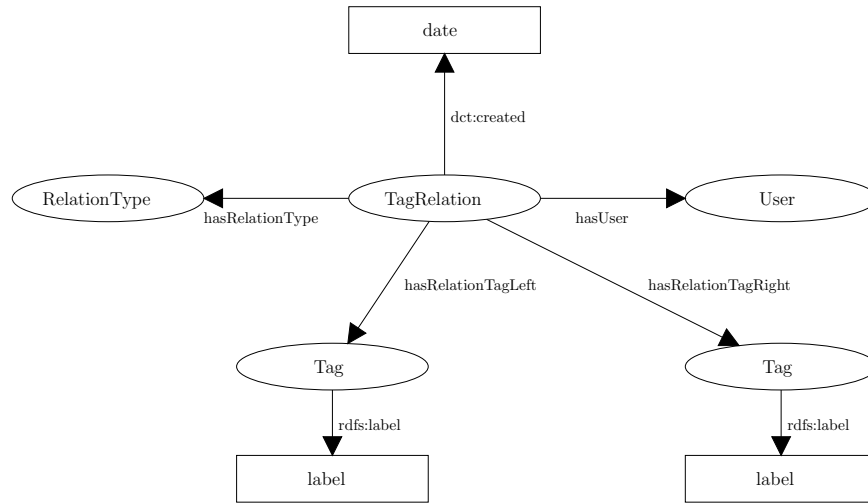


Fig. 4. Tag Relation Model: It is specified that a user defines a relation between two tags.

reified relations. For our case we use the TagRelation class to express an multi-ary relation between a user having stated that two tags are associated by some kind of relation (synonym, broader, narrower, related or ignored) at some point in time. We use RDF for storing this kind of relation graphs. When no namespace is given we use our default namespace. Additionally we reuse the well established vocabulary of Dublin Core¹³ for storing the time the user has created the relation between to tags (`dct:created`). Tags are fully qualified resources and get therefore gets an own URI as identifier¹⁴. The original tag as string is linked to the tag URI by `rdfs:label`, a standard RDF Schema property¹⁵.

Fig. 5 shows an excerpt of a possible resulting thesaurus. There are several tags (e.g. “Java” and “Python”) defined as narrower terms to “Programming”. This information can be utilized when querying for entities/ person that have to do with Programming. An easy example can be: Somebody is looking for a programmer needed for a project. In a standard tagging application searching for “Programming” would not return people tagged with “Java”. Having the tagging application backed by a thesaurus the query can be expanded with defined narrower terms. This basic example for a query extension might also be possible with a better standard thesaurus. When it comes to some narrow domains or

¹³ <http://dublincore.org/documents/dc-rdf/>

¹⁴ <http://www.w3.org/TR/webarch/#uri-benefits>

¹⁵ <http://www.w3.org/TR/rdf-schema/>

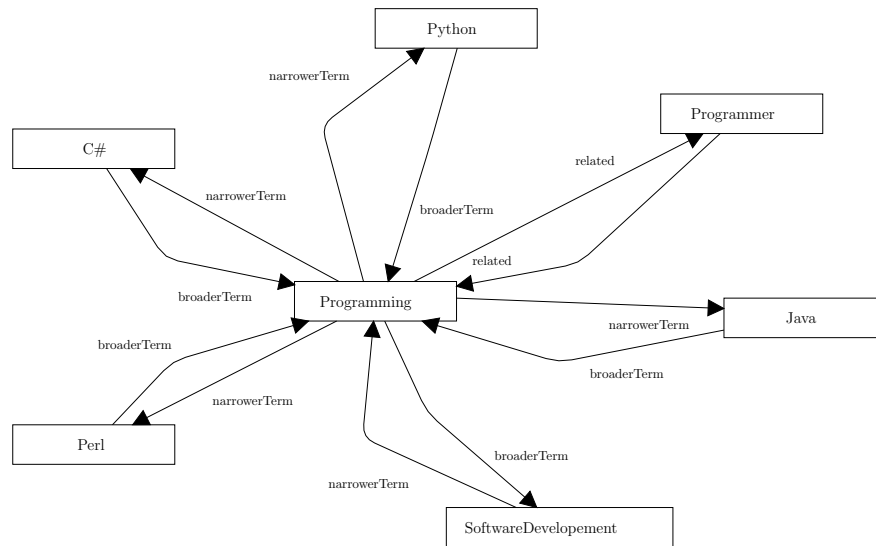


Fig. 5. Example: Relations between tags.

a new terminology emerges (e.g. “cloud computing”, “ajax”, “web2.0”, “grails”, “scala”, etc.) a standard thesaurus can not provide the necessary assistance.

4 The Use Case Competencies and Topics

For the case of the KoPIWA OI platform the previously discussed paradigm can be used for the construction of a thesaurus. The two interesting domains in the context of open innovation are competences and topics. A thesaurus can be utilized to determine semantic similarities between tags for competence and topics respectively. These similarities can e.g. in turn be used to compose teams (by heterogeneity or homogeneity of competencies) or to find items related to similar topics in the open innovation platform.

In section 1.3 we describe an approach for creating a collaborative thesaurus with similar structure as defined in the ISO 2788. For the use case of open innovation in work in progress we built a main infrastructure in the KoPIWA platform that implements the thesaurus.

Thereby, it is possible to: define relations between tags (Related term), group synonymous tags together into so called “tag bags” that represent concepts (Use synonym) and hypernym and hyponym concepts (Broader Term, Narrower Term) by ordering these tag bags into a (multi-)hierarchical structure.

A practical example where a tag thesaurus can be employed is an innovation jam with the topic electric car. People can publish ideas and statements

regarding current and future developments in the area of vehicles driven by electric motors. Depending on the individual background or gusto user can select between different terms having the same meaning (synonym). Typical example tags¹⁶ are “double-layer capacitors”, “supercapacitors”, “pseudocapacitor”, “ultracapacitors”, “electrochemical double layer capacitors”, “EDLC”. They all refer to the same concept¹⁷. An example for different abstraction level can be “on-board energy storage” or “on-board battery packs”. With a backing thesaurus search and navigation can be supported accordingly. Search queries can be either expanded automatically with synonym terms or alternatively a user can be given assistance in Google like style “did you mean ...”. Further a tree/ graph user interface can be provided in which a user can navigate through the set of topics based on the relations defined in the thesaurus. In rather new and innovative fields normally there exist no elaborated thesauri which could be alternatively used. Our proposed thesaurus collaboratively defined by end user is therefore an ideal solution.

4.1 Research Goals

After applying our semi-automatic approach we want to address the following questions:

1. What is the participation pattern of users in the constructions of the tag thesauri?
2. How do our semi-automatically generated tag structures relate to automatically generated ones?
3. As how useful users perceive services that are mainly and directly based on a tag thesaurus?
4. Are the results of functions that model heterogeneity or homogeneity of competencies based on path lengths in the structure of these thesauri (as part of services) in congruence with expectations?
5. Is there a fixed point to which the evolution of the tag structure converges?

To investigate our research question we will use the tag thesaurus editor.

More precisely, we plan to evaluate the first question by observing the generated structures and the quantity of unstructured tags left as “orphans” in the systems. For example how many topic tags used in blog and forum entries, have not been categorized in one of the existing topic tags?

For the second question we want to compare the results of the previously discussed semi-automatic approach for competencies with the results of competence ontologies extracted from job advertisements [34] and analyze the differences.

The research question on homogeneity and heterogeneity as well as the research question on the usefulness of services is addressed through interviews with lead users about the quality of the service results.

The last research question will be investigated through systematic sampling and analyzing the structural dynamics of the thesaurus.

¹⁶ Variations in spelling are possible.

¹⁷ For a description see <http://en.wikipedia.org/wiki/Ultracapacitor>

5 Conclusion and Future Work

We have described a way of creating a thesaurus based on social tagging data. This thesaurus can be used as a base for determining similarities between tags – in the context of open innovation especially competences.

We plan to test combinations of the described approaches in a prototypical implementation of a tagging framework. As data bases we are currently using globally available knowledge management tools of the Siemens AG, namely a wiki and blogging platform. Additionally we are planning to apply the described approach in the KoPIWA Open Innovation community.

We also experimented with data from external system, such as delicious (see [16]) and are planning to evaluate at least unsupervised methods on several other tagging data sources.

6 Acknowledgements

Parts of this paper are based on results of research within the framework of the Theseus project¹⁸, more precisely the Use Case Alexandria. The project was funded by means of the German Federal Ministry of Economy and Technology under the promotional reference “01MQ07012”.

References

1. H. S. Al-Khalifa, H. C. Davis, and L. Gilbert. Creating structure from disorder: Using folksonomies to create semantic metadata. In *the 3rd International Conference on Web Information Systems and Technologies (WEBIST)*, 2007.
2. T. Bocek, E. Hunt, and B. Stiller. Fast Similarity Search in Large Dictionaries. Technical Report ifi-2007.02, Department of Informatics, University of Zurich, April 2007.
3. C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic grounding of tag relatedness in social bookmarking systems. *The Semantic Web - ISWC 2008*, pages 615–631, 2008.
4. H. Chesbrough, W. Vanhaverbeke, and J. West, editors. *Open Innovation: Researching a New Paradigm*. Oxford University Press, USA, 10 2006.
5. M. C. Daconta, L. J. Obrst, and K. T. Smith. *The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management*. Wiley, 5 2003.
6. M. Ester and J. Sander. *Knowledge Discovery in Databases. Techniken und Anwendungen*. Springer, Berlin, 2000.
7. W. Gaus. *Dokumentations- und Ordnungslehre: Theorie und Praxis des Information Retrieval*. Springer, Berlin, 5., überarb. a. edition, 4 2005.
8. S. Golder and B. A. Huberman. The structure of collaborative tagging systems, Aug 2005.
9. M. Grahl, A. Hotho, and G. Stumme. Conceptual clustering of social bookmarking sites. In *7th International Conference on Knowledge Management (I-KNOW '07)*, pages 356–364, Graz, Austria, SEP 2007. Know-Center.

¹⁸ <http://www.theseus-programm.de/home/default.aspx>

10. T. Gruber. Ontology of folksonomy: A mash-up of apples and oranges, November 2005. <http://tomgruber.org/writing/ontology-of-folksonomy.htm>.
11. J. Han and M. Kamber. *Data Mining, Second Edition, Second Edition : Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, 2005.
12. J. Hendler and J. Golbeck. Metcalfe's law, Web 2.0, and the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1):14–20, 2008.
13. A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Emergent semantics in bibsonomy. In C. Hochberger and R. Liskowsky, editors, *GI Jahrestagung (2)*, volume 94 of *LNI*, pages 305–312. GI, 2006.
14. A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Folkrank: A ranking algorithm for folksonomies. In *Proc. FGIR 2006*, 2006.
15. A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information Retrieval in Folksonomies: Search and Ranking. In *The Semantic Web: Research and Applications, 3rd European Semantic Web Conference, ESWC 2006, Budva, Montenegro*, volume 4011 of *Lecture Notes in Computer Science*, pages 411–426, Berlin/ Heidelberg, June 2006. Springer.
16. W. C. Kammergruber, M. Viermetz, and C.-N. Ziegler. Discovering communities of interest in a tagged on-line environment. In *CASoN2009: Proceedings of the 1st International Conference on Computational Aspects of Social Networks*, 2009.
17. B. Krause, C. Schmitz, A. Hotho, and G. Stumme. The anti-social tagger - detecting spam in social bookmarking systems. In *Proc. of the Fourth International Workshop on Adversarial Information Retrieval on the Web*, 2008.
18. D. Laniado, D. Eynard, and M. Colombetti. Using WordNet to turn a folksonomy into a hierarchy of concepts. *Proceedings of SWAP 2007, the 4th Italian Semantic Web Workshop*, page 192, 2007.
19. C. D. Manning, P. Raghavan, and H. Schuetze. *Introduction to Information Retrieval*. Cambridge University Press, July 2008.
20. B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *18th International World Wide Web Conference*, pages 641–641, April 2009.
21. C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPertext '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM Press.
22. A. Mathes. Folksonomies - cooperative classification and communication through shared metadata. *Computer Mediated Communication - LIS590CMC*, December 2004.
23. J. Panyr. Thesauri, Semantische Netze, Frames, Topic Maps, Taxonomien, Ontologien – begriffliche Verwirrung oder konzeptionelle Vielfalt? *Information und Sprache. Festschrift für Harald H. Zimmermann*, pages 139–151, 2006.
24. T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet:: Similarity-measuring the Relatedness of Concepts. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1024–1025. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.
25. I. Peters. *Folksonomies. Indexing and Retrieval in Web 2.0 (Knowledge & Information: Studies in Information Science)*. De Gruyter, 1 edition, 10 2009.
26. M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
27. W. G. Stock and M. Stock. *Wissensrepräsentation: Auswerten und Bereitstellen von Informationen*. Oldenbourg, 5 2008.

28. J. T. Tennis. Social tagging and the next steps for indexing. In J. Furner and J. T. Tennis, editors, *Proceedings 17th SIG/CR Classification Research Workshop*, 2006.
29. E. Tonkin. Searching the long tail: Hidden structure in social tagging. *Proceedings of the 17th SIG Classification Research Workshop*, 2006.
30. Tudhope, Douglas, Binding, Ceri, Blocks, Dorothee, Cunliffe, and Daniel. Query expansion via conceptual distance in thesaurus indexed collections. *Journal of Documentation*, 62(4):509–533, 2006.
31. J. Voss. Tagging, folksonomy & co - renaissance of manual indexing?, Jan 2007.
32. T. V. Wal. Folksonomy. Folksonomy Coinage and Definition, 2007. <http://vanderwal.net/folksonomy.html>.
33. K. Weller. Folksonomies and Ontologies. Two New Players in Indexing and Knowledge Representation. In H. Jezzard, editor, *Applying Web 2.0. Innovation, Impact and Implementation*, pages 108–115, 2007.
34. S. Ziebarth, N. Malzahn, and H. U. Hoppe. Using data mining techniques to support the creation of competence ontologies. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED 2009)*, Brighton, England, July 2009.