Automatic Taxonomy Extraction through Mining Social Networks

Mario Kubek[1], Jürgen Nützel[1], and Frank Zimmermann[1]

4FriendsOnly.com Internet Technologies AG,
98693 Ilmenau, Germany,
{mk, jn, fz}@4fo.de

Abstract. Social networks are a hugely popular part of the Internet. Not only do they offer a convenient way for like-minded people to get in contact with each other, they also represent a valuable knowledge resource. For instance, people publish personal images with own descriptive attributes, or they annotate music while listening to it. These attributes can be of general or rather specific nature, assigned according to the knowledge of the tagging person. Due to the fact that many people annotate the same items with topically related words, often a common agreement regarding their semantic categorization is achieved. This approach is often referred to as building a folksonomy, a portmanteau of the words folk and taxonomy. Unlike this term suggests, a taxonomy cannot be explicitly obtained by simply analyzing the tags of a set of items. This paper presents a method to automatically extract taxonomies from social networks by analyzing co-occurring terms assigned to items and calculating their semantic hierarchy, more precisely by determining the super- and subclasses of terms from a specific tag set based on conditional relative frequencies. Furthermore, practical benefits derived from this method, e.g. to calculate similarities of maybe sparsely or completely differently annotated items, are discussed as well. Additionally, these techniques can also be used to detect inconsistencies among annotations of a specific item.

Keywords: taxonomy extraction, social networks, statistical text mining, co-occurrence analysis, content recommendation, GlobalMusic2one

1 Introduction

In the last years many social platforms have emerged in the Internet. Starting from social bookmarking services like Delicious and Digg over picture and video sharing platforms like Flickr and Youtube to social networks like Facebook and Xing these services and platforms are being used by millions of Internet users day by day. These platforms make it easy for users to share information with each other. Moreover, users are encouraged to categorize and rate it by publishing opinions or annotating content with descriptive attributes, usually called

tags. This collective categorization is commonly referred to as building a folksonomy, a portmanteau of the words folk and taxonomy. Other than this term might suggest, a semantic hierarchy cannot be explicitly retrieved from these annotations and categorizations. However, it is possible to extract taxonomies from these generally unstructured masses of tags by analyzing their semantic relatedness through statistical co-occurrence analysis in order to gain super- and subclasses of terms from a specific tag set. Normally, co-occurrence analysis only yields a single value as a measure of relatedness between two terms. While this approach is sufficient to determine the degree of a semantic relationship of two terms in a textual document, real-world relationships, that are normally of asymmetric nature, cannot be properly acquired by using this approach. Think of the following relation: A BMW vehicle could be associated with a car at 90 percent, but a car can be associated with a BMW vehicle at only 15 percent. Such asymmetric relations can be easily gained by calculating the conditional probability as a co-occurrence measure that a term A occurs under the condition that another term B occurs in close range to term A. This way more general terms can be separated from specific terms, which is a prerequisite to build a taxonomy. This paper presents a method to extract taxonomies through mining social networks based on calculating the aforementioned asymmetric term relations. Furthermore, practical backgrounds and benefits of this method are discussed as well. Based on such semantic term hierarchies, it is possible to calculate similarities of maybe sparsely or completely differently annotated items. Another field of application can be seen in the automatic enrichment of sparsely annotated content with terms from a taxonomy. Additionally, the application of such taxonomies in Internet portals can also enhance the browsing experience for users, as they could easily navigate to more general or to more specific terms depending on their information needs. The next section explains the basics of statistical co-occurrence analysis. Section three focuses on the methods to extract taxonomies from social networks. As an example the research project GlobalMusic2one applies these techniques to calculate similarities and hierarchies of music annotations. In section four a taxonomy of music tags calculated from a Last.fm dataset based on the methods from section three is introduced. Section five concludes the paper and gives a prospect on future developments regarding automatic taxonomy extraction in the Internet.

2 Detecting Semantic Relations

The occurrence of two words in a text section next to each other is called co-occurrence or syntagmatic relation[1]. Co-occurrences that appear above chance are called significant co-occurrences. They can be used to detect semantic relationships among words. The most prominent kinds of co-occurrences are words that occur as immediate neighbours and words that occur together in a sentence. In the following considerations we will focus on the latter ones. There are several well-established measures to calculate the statistical significance of such word pairs by assigning them a significance value. If this value is above a

preset threshold the co-occurrence can be regarded as significant and a semantic relation between the involved words can often be derived from it. Rather simple co-occurrence measures are for instance the frequency count of co-occurring words and the similar Dice and Jaccard coefficients[2]. More advanced formulas rely on the expectation that two words are statistical independent which is a usually inadequate hypothesis. They then calculate the deviation from their observation of real corpus data to their expectation. A significant deviation leads therefore to a high co-occurrence value. For this purpose the number of sentences in which the two words of interest A and B occur separately, the number of sentences in which these words occur together and the number of all sentences in the corpus are taken into account. Co-occurrence measures based on this hypothesis are for instance the mutual information measure[2], the poisson collocation measure[3] and the log-likelihood ratio[4]. Stimulus-response experiments show that co-occurrences found to be significant by these measures correlate well with word associations by humans. The calculation of sentence co-occurrence values usually yields a symmetric relation of a word A with B and vice versa. This symmetry is of practical benefit when it comes to visualizing word associations on a map [5][6] using co-occurrence matrices or providing solutions for query expansion[7], but it is not a proper basis to build up hierarchies of real-world items, as their semantic relations are normally of asymmetric nature. In [8] a first approach to calculate word hierarchies using co-occurrence analyzation is discussed. Conditional probabilities from the probability theory can provide a solution to determine asymmetric relations of words in a text corpus. Generally, a conditional probability is the probability of an event A under the condition of another event B and is defined by this formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

With reference to statistical text mining, the conditional probability that a specific word A occurs under the condition that another word B co-occurs with word A on sentence level can be used as an asymmetric co-occurrence measure and can also be seen as a measure of the degree of the association of word B with word A. It can yield different values for the association of a word A with a word B and for the association of the same word B with the same word A. Based on this asymmetry it is possible to calculate the aforementioned taxonomies. For this purpose it is also necessary to correctly consider the cardinality of term sets and how they relate to each other. Therefore, conditional relative frequencies that correspond well to these conditional probabilities can be easily calculated when dealing with large datasets. This way it is not necessary to calculate the probabilities themselves but to just rely on the cardinality of term sets which are much easier to determine. The next section describes a method to apply this approach to extract taxonomies from social networks that use social tagging to categorize items.

3 Taxonomy Extraction

Before we introduce this method, some basic definitions need to be given to understand the terminology used. Term B is a subclass of a term A, if the set of items tagged with B is a proper subset of the set of items tagged with A and the set of items tagged with A is not a proper subset of the set of items tagged with B. In this case term B can be absolutely associated with term A, but term A cannot be absolutely associated with term B. Term A itself is then a superclass of term B. If a term A is a superclass of term B, then term B cannot be a superclass of term A. Additionally, it is possible, that term B itself is a superclass of another term C. Term A is then also a superclass of term C. In this case we speak of a transitive relation. If a term A has more than one subclass, then these subclasses are called cohyponyms. They differ in at least one semantic feature. Hence, these relations of super- and subclasses are asymmetric, which is a prerequisite to build up term hierarchies.

3.1 The Algorithm

Now we want to describe the basic algorithm to automatically build a taxonomy of tag sets from social networks. The thresholds used are empirically found values.

1. Acquire all tags of a (preselected) set of items (e.g. images or songs) in a social network, whereby co-occurring tags (tags assigned to one item) must be saved in a data structure that separates the tags of different items e.g. a hash map with item identifiers as keys and Lists of tags as values.
2. Calculate the association $Assn$ of tag A with tag B for all tags. For this purpose the following parameters need to be determined:
   - the number of items annotated with tag A: $|A|$
   - the number of items annotated with tag B: $|B|$
   - the number of items annotated with both tags: $|A \cap B|$

   This association can be determined by the following formula which is the conditional relative frequency of tag B for the items annotated with tag A:

$$Assn(A \rightarrow B) = \frac{|A \cap B|}{|A|} \tag{2}$$

   The values calculated by this formula lie within the range of 0 and 1, whereby values near 0 point to an unimportant semantic relation and values near 1 indicate a strong association. However, to determine the real value of a semantic relation of a tag A with another tag B one has to consider the association of tag A with tag B and the association of tag B with tag A.
3. Determine a direct superclass A of a tag B by applying the following rules:
   - Find out all superclasses of tag B:
     - the association of tag B with another tag must be greater than the association of this other tag with tag B

    - the association of tag B with this other tag must be greater than 0.80
    - the association of this other tag with tag B must be less than 0.50 and greater than 0.09 to filter out synonyms and less important tags
  - Remove all indirect superclasses (superclasses of superclasses):
    - order all superclasses of tag B according to their association with B
    - determine indirect superclasses of B by comparing all superclasses of tag B with each other and selecting those superclasses that have an association with another superclass of this set of at least 0.80
    - rule out these indirect superclasses
4. Determine direct subclasses of a tag B by applying the following rules:
  - Find out all subclasses of tag B:
    - the association of another tag with tag B must be greater than the association of tag B with this other tag
    - the association of this other tag with tag B must be greater than 0.80
    - the association of tag B with this other tag must be less than 0.50 and greater than 0.09 to filter out synonyms and less important tags
  - Remove all indirect subclasses (subclasses of subclasses):
    - order all subclasses of tag B according to B's association with them
    - determine indirect superclasses of B by comparing all subclasses of tag B with each other and selecting those subclasses that have an association with another subclass of this set of at least 0.80
    - rule out these indirect superclasses

3.2 Discussion

The result of this algorithm is an acyclic graph of hierarchically connected tags as nodes. This method works because people assign general and specific tags to items at the same time. This of course depends on their knowledge of the items to be annotated. However, it is a sensible idea to clean up tags sets prior to taxonomy extraction by removing stop words, by applying methods of base form reduction like stemming or lemmatization or by selecting specific word forms like nouns via part-of-speech tagging or dictionary lookups. Also the removal of rarely used tags and the usage of tags that are likely to be important by considering their word length in combination with their weighted frequency in the tag set can significantly improve the quality of the resulting taxonomy. This approach has already been implemented in the Firefox extension "FireMatcher"[9] to automatically generate search queries consisting of catchwords of analyzed documents and even works without consulting a reference corpus.

4 A Taxonomy of Last.fm tags

We evaluated this taxonomy extraction algorithm by analyzing music tags in the social music service Last.fm. In this process, 1.2 million relations of 2800

tags of the 1500 most popular songs of 2009 in Germany have been calculated, representing a high diversity of music styles. To enhance the quality of the resulting taxonomy we ruled out tags that have been assigned less than 5 times in addition to the rules in the previous section. Figure 1 depicts only a small fraction of the complete taxonomy.

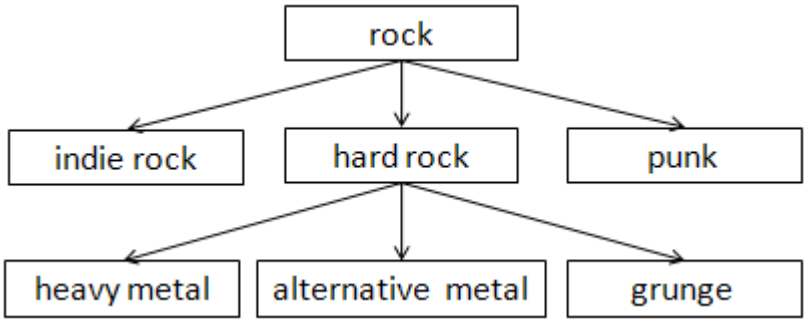

Fig. 1. Small part of the entire taxonomy calculated from Last.fm tags

The tag associations that lead to this taxonomy are presented in the following table in form of a matrix and have been calculated during step 2 of the algorithm in the previous section. Logically, the diagonal of this matrix is left empty because there is no sensible association between identical tags.

Association of with	rock	hard rock	indie rock	punk	heavy metal	alternative metal	grunge
rock		0.46	0.45	0.43	0.16	0.18	0.19
hard rock	1.0		0.36	0.53	0.32	0.36	0.34
indie rock	0.99	0.37		0.41	0.06	0.10	0.19
punk	0.98	0.56	0.01		0.16	0.16	0.28
heavy metal	0.95	0.87	0.17	0.41		0.62	0.26
alternative metal	0.96	0.90	0.26	0.38	0.58		0.22
grunge	1.0	0.83	0.45	0.64	0.23	0.22	

Table 1. Calculated associations between the tags of the taxonomy in figure 1

Based on the results obtained from these calculations, it was found that this method is a proper way to determine class relations and class hierarchies in the hybrid music recommender system GlobalMusic2one[10], which users "train" by adding new musical categories or classes and assigning them example songs.

Moreover, by using class hierarchies users will be able to navigate to more specific or to rather general classes using hierarchical tag cloud navigation. Depending on particular user queries related super- and subclasses could be suggested as expansion terms or separate queries to broaden or to constrain the search space. Based on such semantic term hierarchies it is also possible to calculate similarities of maybe sparsely or completely differently annotated items when their annotations are mapped onto such pre-calculated taxonomies and paths of their subtrees are compared, e.g. using the vector space model. This technique could also be used to detect inconsistencies among annotations of a specific item. This might be the case when its annotations belong to different subtrees of the used taxonomy.

5 Conclusion

In this paper we introduced a method to calculate taxonomies from annotations in social networks based on statistical co-occurrence analysis. We described the prerequisites to apply this method, discussed practical benefits of this approach and presented an example taxonomy of Last.fm tags. We also gave some hints on how to enhance the quality of the resulting taxonomy. As mentioned in the previous section, such taxonomies calculated from collaboratively acquired tags can be the basis for new techniques to determine content relations. This also opens the possibility for new content recommendation algorithms. Users who like content annotated with specific tags might also be interested in content annotated with more general tags or with their cohyponyms. These classes of tags could also be the subject of further interesting derivative research questions. Cohyponyms differ in at least one semantic feature. Therefore, it might be interesting to find out if it is always possible to technically confirm these presumed semantic differences e.g. with the help of digital signal processing by determining discriminating features? This might be an immanent question when it comes to address the problem of differences between computed content similarities and human perception ("semantic gap"). Furthermore, it is conceivable to use such taxonomies as some kind of reference corpora to automatically assign tags from a taxonomy to content based on extracted content features. This of course will only work if a mapping from these features to the terms in the taxonomy can be applied. These examples show that there are numerous ways to make use of detected semantic relations of annotations in social networks and that it is sensible to transform them into a taxonomy which does not only organize them into a superclass-subclass hierarchy but could also pave the way for derived use cases.

Acknowledgments

The authors would like to thank Dr. Thomas Böhme of the Institute of Mathematics at the Technical University Ilmenau for his valuable directions and advices

regarding the interpretation of the semantic relations found during the taxonomy computations.

References

1. Heyer, G., Quasthoff, U., Wittig, Th.: Text Mining – Wissensrohstoff Text, W3L Verlag Bochum, (2006)
2. Buechler, M.: Flexibles Berechnen von Kookkurrenzen auf strukturierten und unstrukturierten Daten, Master's thesis, University of Leipzig (2006)
3. Quasthoff, U., Wolff, Chr.: The Poisson Collocation Measure and its Applications, In: Proc. Second International Workshop on Computational Approaches to Collocations, Wien (2002)
4. Dunning, T.: Accurate methods for the statistics of surprise and coincidence, Computational Linguistics, 19(1):61–74 (1994)
5. Richter, M: Analysis and Visualization for Daily Newspaper Corpora, Proceedings of RANLP (2005)
6. Kubek, M., Nützel J.:, Novel Interactive Music Search Techniques, In: Proc. of the 7th International Workshop for Technical, Economic and Legal Aspects of Business Models for Virtual Goods incorporating the 5th International ODRL Workshop (VG'09), Nancy (2009)
7. Kubek, M., Witschel H.F.: A System for Automatic Query Expansion in a Browser-based Environment, In: Proc. of the 6th International Workshop for Technical, Economic and Legal Aspects of Business Models for Virtual Goods (VG'08), Poznań (2008)
8. Morita, K et al.: Word classification and hierarchy using co-occurrence word information. Information Processing and Management, 40(6): 957-972 (2004)
9. Website of FireMatcher, http://www.firematcher.com/
10. Website of GlobalMusic2one, http://www.globalmusic2one.net/