

On the Distinctiveness of Tags in Collaborative Tagging Systems

André Gohr
Computer Science
Martin-Luther-University
Halle, Germany
gohr@informatik.uni-halle.de

Myra Spiliopoulou
Computer Science
Otto-von-Guericke-University
Magdeburg, Germany
myra@iti.cs.uni-magdeburg.de

Alexander Hinneburg
Computer Science
Martin-Luther-University
Halle, Germany
hinneburg@informatik.uni-halle.de

Ricardo Usbeck
Computer Science
Martin-Luther-University
Halle, Germany
usbeck@informatik.uni-halle.de

ABSTRACT

We study in a quantitative way whether the most popular tags in a collaborative tagging system are distinctive features when looking at the underlying content. For any set of annotations being helpful in searching, this property must necessarily hold to a strong degree. Our initial experiments show that the most frequent tags in CiteULike are distinctive features, despite the process of annotating documents is not centrally coordinated nor correction mechanisms like in a Wiki-system are used.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

General Terms

Experimentation, Measurement

Keywords

Text Mining, Collaborative Information Systems, Tagging

1. INTRODUCTION

The Semantic Web is a vision that wants to bring order to the unstructured web. Unstructured resources like text, images or videos are annotated in machine readable form. The annotations of resources that semantically fit together are either identical or could be logically deduced using ontologies. The main goal is to make semantic relatedness that is recognized by humans amenable to logic operations that could be also run on computer. This helps searching the semantic web and has the twofold benefit of (i) distributing

the annotation costs among many individuals and (ii) capturing many different semantic aspects. Beside those nice things, however, the unrestricted use of annotations called tags may undermine the very goals of semantic web when tags are used in a wrong way. It is often claimed that the social processes at work in a collaborative tagging system prevent such undesired effects. However, this is an open question. Here, we study in a quantitative way whether the most popular tags in a collaborative tagging system are distinctive features when looking at the underlying content. For any set of annotations being helpful in searching, this property must necessarily hold to a strong degree.

Related works that studied the tagging process include: Why do people tag [4], usage patterns of tags [1] and discovering leaders in tag usage [2]. These studies take the tag names as their face values and/or look at the co-occurrences between users, tags and resources only. Here, we treat tags as IDs and do not rely on the particular strings chosen. Furthermore, we concentrate on the particular case of resources being text documents. When relating tags to each other, the word distributions of underlying documents are used. This goes beyond simple co-occurrence analysis between tags and resource IDs. Thus, our analysis detects semantically related tags even when the particular tags do not frequently co-occur.

To study whether the most popular tags are distinctive features, we repeat the following experiment many times. We simulate a virtual user who wrongly annotates documents with a particular tag, which we call target tag. The documents this user annotates with the target tag are randomly selected from the documents annotated with another tag called source tag. When target tag and source tag are distinctive features, the documents of the virtual user should significantly differ from those documents that real users annotated with the target tag. We test this by checking whether the virtual user is a distance-based outlier with respect to the real users. The virtual user is an outlier, when he ranks top among the users who annotated with the target tag with respect to descending distance. The experiment is repeated many times drawing source and target tags randomly from the most frequent tags in the system. When those tags are distinctive features, the simulated virtual user

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WIMS '11, May 25-27, 2011 Sogndal, Norway

Copyright 2011 ACM 978-1-4503-0148-0/11/05 ...\$10.00.

should be detected as an outlier many times.

The outcome of the experiment depends on the chosen distance between the subset of documents annotated by a user and the overall set of documents under the target tag. We compute this distance using Kullback-Leibler (KL) divergence between the multinomial word distributions of the overall set of documents under the target tag and the individual documents of the particular user. To gain more insights into the results of the experiment, we study the influence of sparseness of the word distributions onto the KL-divergence.

As not all words of the vocabulary appear in a document, the estimated word distribution of the document assigns zero probabilities to the not included words. That sparseness influences the distance measure used to compare the documents annotated by a user with the overall set of documents under the target tag. We demonstrate and quantify the influence.

Furthermore, we study whether the sparseness is important for the distinctiveness of tags. We use a topic model (Probabilistic Latent Semantic Analysis [3]) to reduce the dimensionality of the word distributions from the size of the vocabulary to a much smaller number K . The reduced distributions represent the documents in a low-dimensional space. The newly computed distributions are not sparse. We redo the above described experiments on the new representations in order to check whether the initial results on word distributions depend on the sparseness of the original document representations.

In the following, we describe in Section 2 the setup of the experiments using data from the CiteULike platform. In Section 3, we report results of a sanity check experiment using Bible documents and a run on the real data from CiteULike. Last, we discuss initial results and explain future work.

2. METHODS AND MATERIALS

We study the distinctiveness of frequent tags in data from CiteULike.org¹, which is a platform that allows to collaboratively annotate scientific articles. These data comprise triples (userID, docID, tag) for each tagging event. That is a user has annotated a document with a tag. We complemented the tagging data with the abstracts corresponding to the document IDs found in www.ncbi.nlm.nih.gov/pubmed (240.000 abstracts), archivx.org (9.000 abstracts) and cite-seerx.ist.psu.edu (13.000 abstracts).

We concentrate on the most frequently used tags with which at least 500 documents (abstract) have been annotated. Neglecting the two top tags *no-tag* and *review*, we are left with a set T of 91 tags, e.g. *protein*, *quantum*, *analysis*, or *imaging*.²

2.1 Tag Distinctiveness

We test the distinctiveness of the most frequent tags T by the following experiment. For randomly chosen target tag t and source tag t' , we construct a set of documents D consisting of documents annotated by real users with t and k documents annotated with t' , which simulates the virtual user. To make the results comparable over the different choices for t and t' , we require the total size of set $|D| =$

¹downloaded March 2010

²Complete list of tags is attached to the PDF file. 

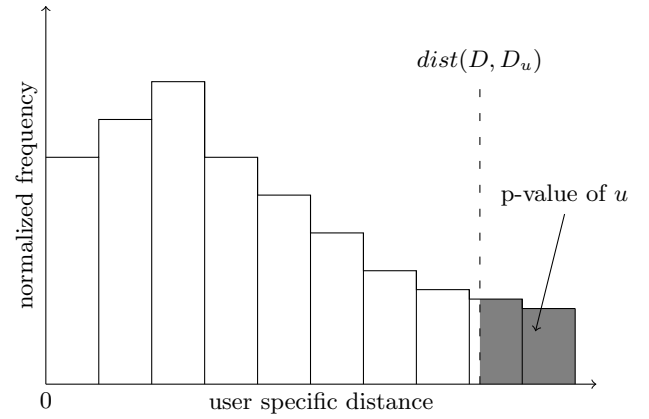


Figure 1: Normalization of user specific distances to p-values. The integral of the whole histogram is normalized to one. The p-value of u is the integral from $dist(D, D_u)$ to ∞ .

500. This particular size is chosen because each tag in T annotates at least 500 documents. The set D includes 500– k documents randomly sampled from the documents real users annotated with t . The k documents of the simulated virtual user u' are randomly sampled from the documents annotated with t' . The set $D_u \subset D$ is the set of documents user u annotated with t . The user u might be a real user or the simulated virtual user.

The target and source tag would be distinct when the documents annotated by the virtual user u' are very different from the documents in D annotated by the other users. Therefore, we define a distance between the documents of a user D_u and the total set D . Let be m the multinomial probability distribution estimated from the word frequencies in D . Analogously, m_d denotes the distribution estimated from the word frequencies in a single document d . The distance between D_u and D is the average Kullback-Leibler (KL)-divergence between the overall word distribution m and the document specific distributions m_d with $d \in D_u$

$$dist(D, D_u) = \frac{1}{|D_u|} \sum_{d \in D_u} KL(m || m_d)$$

To compare results over different choices for target and source tag, the user specific distances are normalized to p-values. The user specific p-value p_u is the percentage of users who contributed to D and have a larger distance as the user u . An explanation of the normalization to p-values is shown in Figure 1. When a user is an outlier, the corresponding p-value p_u should be close to zero meaning that none or only a few users have larger distance.

To ensure comparability between different selections of source and target tag, we used a fixed number of bootstrap users for the p-value normalization that are generated from real users. The documents of a bootstrap user $\hat{u}$ are drawn with replacement from the documents of a randomly selected real user. The distances of the generated document sets $D_{\hat{u}}$ are the basis for the p-value normalization.

The experiment is repeated many times with random choices for t and t' . In each experiment, one virtual user is simulated. When the tags in T are distinctive features, the simulated virtual users should be top ranked with respect to

ascending p-value over the computed p-values of all users. Virtual users with large p-value indicate pairs of source and target tags that are difficult to tell apart when looking at the underlying documents.

2.2 Influence of Sparseness

The outcome of the experiment is influenced by the distance measure used to detect the virtual user as an outlier. We investigate the effects of the sparseness in the representations of the documents onto the chosen distance measure: KL-divergence. Let be W the vocabulary of the documents in D , KL-divergence is defined as

$$\text{KL}(m||m_d) = - \sum_{w \in W} m(w) \log \frac{m_d(w)}{m(w)}$$

The KL-divergence can be interpreted as the average number of bits per word that are wasted, when a message with word distribution m is sub-optimally coded with a code constructed for another word distribution m_d .

When a particular word w does not appear in document d the distribution $m_d(w)$ would be zero. Those words are responsible for the sparseness in the text data. The zero probability of w in m_d would lead to numerical problems when computing KL-divergence as $\log 0$ is not defined. The standard way to avoid the problem is to add a small positive constant to all outputs of the distribution m_d and re-normalize it. Thus, the distribution assigns to word w a small constant $m_d(w) = \epsilon > 0$.

To study the effect of words that are not part of document d , the sum over the vocabulary in the formula of the KL-divergence is split into two parts: P_d^+ consists of summands with words that appear in d and P_d^- consists of summands with words that do not appear in d :

$$\begin{aligned} \text{KL}(m||m_d) &= P_d^+ + P_d^- \text{ with} \\ P_d^+ &= \sum_{w \in W \text{ and } w \in d} -m(w) \log \frac{m_d(w)}{m(w)} \\ P_d^- &= \sum_{w \in W \text{ and } w \notin d} -m(w) \log \frac{m_d(w)}{m(w)} \end{aligned}$$

Note that while the KL-divergence is always non-negative, the parts P_d^+ and P_d^- are not necessarily larger zero. In fact, as the probability mass in m_d needs to be distributed over much less words as in m , we expect for words w appearing in document d that $m_d(w) > m(w)$. Thus, the summands that are part of P_d^+ should be mainly negative and consequently we expect $P_d^+ < 0$. Vice versa, for words w not appearing in d , we have $m_d(w) < m(w)$ and therefore the summands of P_d^- should be positive.

This intuition goes nicely along with the interpretation of the KL-divergence as the average number of bits wasted by using a sub-optimal code wrt. m_d instead of m . The words appearing in document d are overrepresented in m_d compared to m . Thus, these words get shorter codes and some bits per word are saved (negative waste). The words not appearing in d are underrepresented in m_d compared to m . Thus, these words get longer codes and bits per word are wasted (positive waste). The non-negativity of the KL divergence tells us that there are always less bits saved than wasted.

To shed light on the influence of the sparseness of the data onto the quantified distinctiveness of the frequent tags,

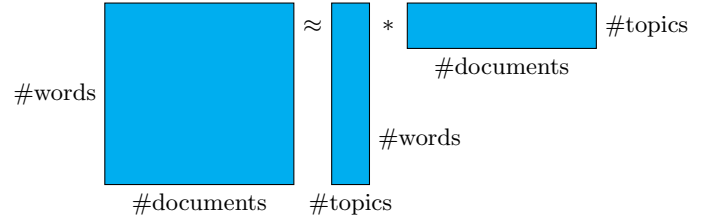


Figure 2: Matrix Decomposition

average values of the parts P_d^+ and P_d^- are reported for real users as well as virtual users. This allows us to tell, whether virtual users differ from real users because of the words virtual users are using or because of words they are not using.

2.3 Dimensionality Reduction

Beside the chosen distance measure between the subset of documents annotated by a user and overall set of documents, the representation of the document themselves influences the outcome the experiments. So far, documents have been represented by multinomial distributions of the vocabulary. The parameters of those distributions are sparse, high dimensional vectors. The question here is: Does the sparseness in the document representations contribute to the distinctiveness of tags? Therefore, we redo the experiments on non-sparse document representations.

To get rid of the sparseness, the document vectors are approximated by low-dimensional vectors. The needed transformation can be thought of as matrix decomposition (Figure 2). The large but sparse word-document matrix, the columns of which are the parameters of the document specific word distributions, is approximated by a product of two much smaller matrices that represent the parameters of the used topic model. The columns of the topic-document matrix are the new, low dimensional document representations. The dimensionality of the original vectors is reduced from the size of the vocabulary to the number of topics.

We compute the reduced representation using the PLSA topic model. The PLSA model approximates the word distribution of a document by a linear combination of word distributions called topics:

$$p(w|d) = \sum_{k=1}^K p(z=k|d) \cdot p(w|z=k)$$

All distributions are multinomial distributions. The variable z is an indicator and probabilistically assigns a particular word w in document d to the topics. The parameters of the distributions are learned by an expectation maximization algorithm that maximizes the likelihood of the set of words in the document set D of the particular target tag at hand. The number of topics K is a given parameter that has to be chosen beforehand. The details of the PLSA model and the EM algorithm can be found elsewhere [3].

The new reduced document representations are the K parameters of the multinomial distribution $p(z|d)$ that is specific for document d . The experiment on tag distinctiveness described in Section 2.1 is redone using the new reduced document representations constructed by PLSA for the document sets of the respective target tags. The comparison of the outcomes of both experiments (original and reduced

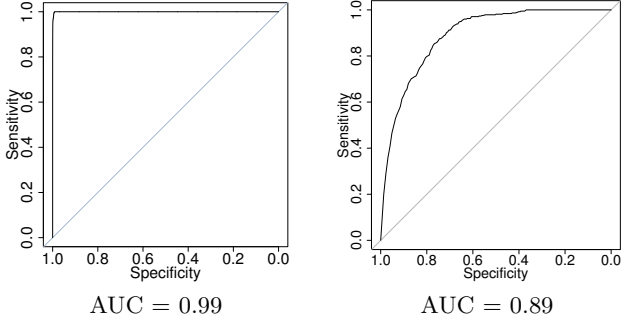


Figure 3: Left: ROC curve with area-under-curve (AUC) value for the experiments with target tags and the *bible* source tag, Right: ROC curve with AUC value for 100 randomly chosen pairs of target and source tag.

document representation) will quantify the contribution of sparseness to tag distinctiveness.

3. RESULTS

Result of the experiments on tag distinctiveness, influence of sparseness and dimensionality reduction are reported.

3.1 Tag Distinctiveness

First, we test whether pairs of tags that are used for different subjects are indeed identified. Therefore, we added the tag *bible*, which subsumes short excerpts from the Old Testament. We choose all tags from T as target tag, respectively. For each target tag t , we used the *bible*-tag as source tag to simulate a virtual user. The experiment is repeated with different numbers of documents associated to the virtual user $k \in \{5, 10, 15, 20, 25, 30\}$. Figure 3, left shows the ROC curve that summarizes the concentration of virtual users at the top in the overall ranking of users with respect to ascending p-values. A ROC curve plots specificity (true negative rate) versus sensitivity (true positive rate). The sharp rise of the curve indicates a large concentration of virtual users at the top positions of the ranking. Also the AUC value of 0.99 indicates a nearly perfect detection of the virtual users as outliers. Thus, as expected, our method shows that the *bible*-tag is very distinct from the rest of the tags.

In the next experiment, 100 pairs of target and source tags are randomly selected from T . Figure 3, right shows the corresponding ROC curve. Here, the detection of the virtual users is not that perfect as in case of the tag *bible*. However, the ROC curve still indicates a large concentration of virtual users at the top positions of the p-value ranking. Also the AUC value of 0.89 is far away from 0.5, which would indicate no concentration of virtual users at the top positions.

The pair of target and source tag with the largest p-value among the 100 randomly chosen pairs is (*gene*, *rmai*). Both tags annotate publications in biological journals. Such related tags among the 100 most frequent tags are responsible for the decrease of the AUC value in this experiment.

3.2 Influence of Sparseness

Words that do not appear in a document cause many zero frequencies in the word distributions of documents, a

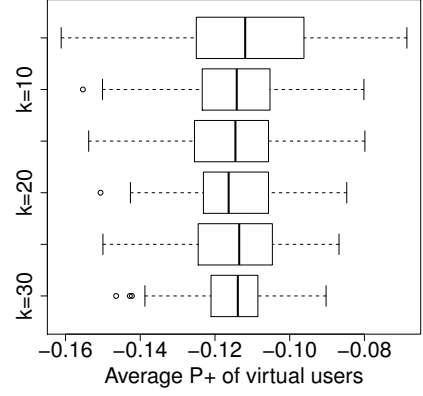


Figure 4: Dependence of Average P^+ on the size k of the virtual user.

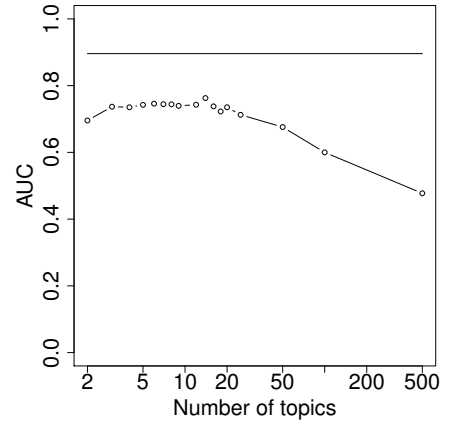


Figure 5: Results of tag distinctiveness experiment on document representations with reduced dimensionality for different number of topics. The horizontal line indicates AUC of the original document representation.

phenomenon called sparseness. The distance between the documents D_u annotated by a user and the overall set of documents D under the target tag averages over the KL-divergence between the word distributions of documents in D_u and the global word distribution of D . For a single document, the KL-divergence is split into a part P_d^+ contributed by words appearing in d and another part P_d^- contributed by words that do not appear in d . The quantity P_d^- is the contribution of the sparse part of d to the KL-divergence of that document. We average both P_d^+ and P_d^- over the documents $d \in D_u$ of a user. These means are averaged again over the real and virtual users respectively. In the end, given a target tag with virtual user, we get two tuples each consisting of average P^+ and average P_d^- : one tuple for real users and the other one for the virtual user.

Figure 6 shows those tuples as points for different sizes of the virtual user for all target tags used in the previous experiment. The plots show clearly that the virtual user has larger Average P^+ than real users, which means according to the interpretation of KL-divergence that the virtual

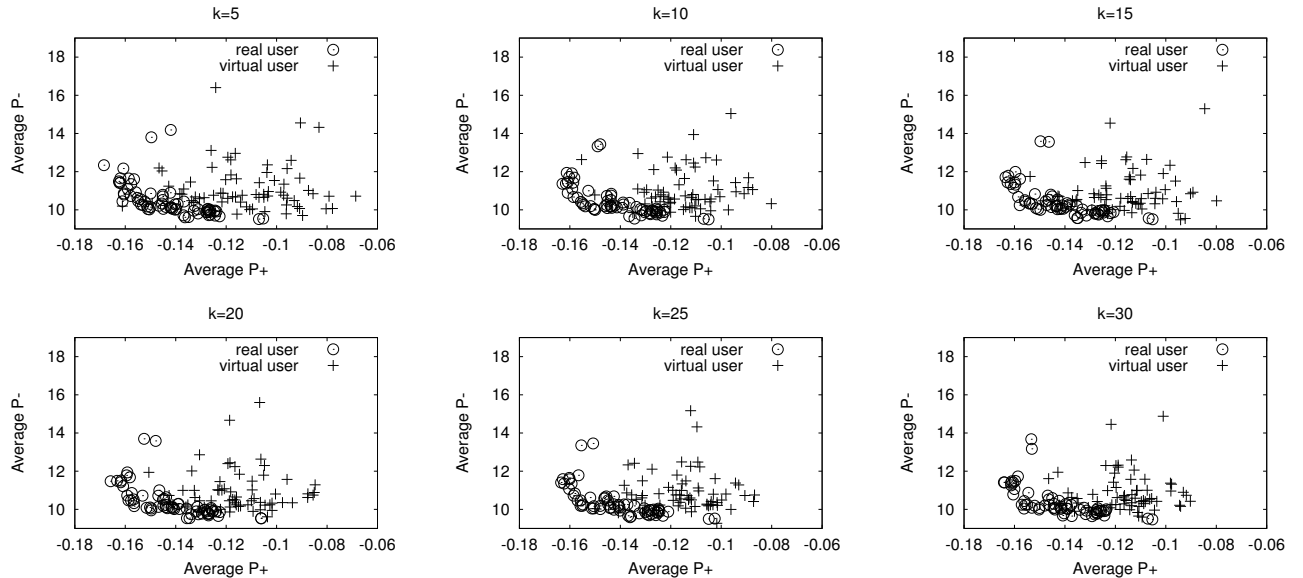


Figure 6: Average P^+ and Average P^- for real and virtual users for all sampled pairs of source and target tag.

user can save less bits per word on average as real users. Furthermore, the plots show that the virtual user has larger Average P^- than real users. This means, that the virtual user wastes more bits on words that do not appear in the respective documents than real users do. Both effects together lead to the result that virtual users have typically larger KL-divergence than real users and could be detected as outliers. Looking at the magnitudes of Average P^+ and Average P^- , we see that Average P^- is two magnitudes larger than Average P^+ . This leads to the conclusion, that the words that do not appear in documents of the virtual user play a more important role for making the virtual user an outlier than the words that do appear in those documents. This suggests that a source and target tag are distinct because the documents under the source tag do not use certain words present in the documents under the target tag.

The size of the virtual user influences mainly the standard deviation of Average P^+ of the virtual user, which is shrinking with growing size k . Figure 4 shows that effect in box-plots for different k of Average P^+ of the virtual user.

3.3 Dimensionality Reduction

In the last experiment, we investigate the influence of the sparseness of the document representations onto the results of the tag distinctiveness experiment. To this end, we eliminated the sparseness from the document representations by reducing the dimensionality of the document representation with Probabilistic Latent Semantic Indexing (PLSA). The new document representations consists of coefficients for topic distributions learned by an EM-algorithm. The number of topics to be learned is an parameter.

We varied the number of topics (reduced dimensionality) and redid the tag distinctiveness experiment to compute ROC curves and AUC values. As result, we got an AUC value for each choice of the number of topics. Figure 5 plots number of topics versus AUC value. It shows that no document representation with reduced dimensionality reached an

AUC value as high as the original document representation with sparseness. This indicates that the document representations learned by PLSA topic models do not capture all aspects of tag distinctiveness.

4. CONCLUSIONS

Our initial experiments show that the most frequent tags in CiteULike are distinctive features, despite the process of annotating documents is not centrally coordinated nor correction mechanisms like in a Wiki-system are used. Thus, those tags should be helpful features in semantic search.

We analyzed further that the distinctiveness between two tags is mainly due to the fact, that words that are present in documents under a one tag are missing in documents under the other tag. Further experiments showed that when sparseness in the document representations is eliminated using topic models some aspects of tag distinctiveness are lost.

In future work, we want to expand the number of analyzed tag pairs. Furthermore, we plan to investigate how the distinctiveness of tags develops over time.

5. REFERENCES

- [1] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32:198–208, April 2006.
- [2] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Discovering leaders from community actions. In *CIKM'08*, pages 499–508, Napa Valley, CA, USA, Oct. 2008. ACM.
- [3] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.
- [4] O. Nov and C. Ye. Why do people tag?: motivations for photo tagging. *Commun. ACM*, 53:128–131, July 2010.