Using the Overlapping Community Structure of a Network of Tags to Improve Text Clustering

Nuno Cravino
nuno.cravino@dcc.fc.up.pt

José Devezas
jld@dcc.fc.up.pt

Álvaro Figueira
arf@dcc.fc.up.pt

CRACS/INESC TEC, Faculdade de Ciências, Universidade do Porto
Rua do Campo Alegre, 1021/1055, 4169-007 Porto, Portugal

ABSTRACT

Breadcrumbs is a folksonomy of news clips, where users can aggregate fragments of text taken from online news. Besides the textual content, each news clip contains a set of metadata fields associated with it. User-defined tags are one of the most important of those information fields. Based on a small data set of news clips, we build a network of co-occurrence of tags in news clips, and use it to improve text clustering. We do this by defining a weighted cosine similarity proximity measure that takes into account both the clip vectors and the tag vectors. The tag weight is computed using the related tags that are present in the discovered community. We then use the resulting vectors together with the new distance metric, which allows us to identify socially biased document clusters. Our study indicates that using the structural features of the network of tags leads to a positive impact in the clustering process.

Categories and Subject Descriptors

G.2.2 [**Discrete Mathematics**]: Graph Theory—*graph algorithms, network problems*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*; I.5.3 [**Pattern Recognition**]: Clustering—*algorithms, similarity measures*

General Terms

Algorithms, Experimentation

Keywords

Text clustering, user-defined tags, network of co-occurrence of tags, overlapping community structure, news clips

1. INTRODUCTION

Tagging is a frequent behavior of people consuming online information. Many well-known online systems take advantage of tags in order to improve the organization of the stored content, or to enhance the accuracy of search results from user queries. Collaborative tagging systems [4] such as Delicious and Flickr (to name a few) have had great popularity recently. These systems allow their users to tag web pages, or photos, eventually in a collaborative way. This folksonomic [4] way of creating tags, while associating them to web content, is helping systems to automatically enhance the existing clustering processes and grouping techniques. We believe our new classification technique has potential to produce even better results by making it closer to a human-only classification, while being performed automatically.

While collaborative tagging offers many advantages over the use of controlled vocabularies [11], they also suffer from several limitations at the same time due to the unrestricted nature of tagging [4]. The fact that many tags are ambiguous has limited the effectiveness of collaborative tagging systems in document description and retrieval. Unlike keywords, hierarchies or even taxonomies, tags usually lack any form of explicit organization and normalization.

The Breadcrumbs system [1] helps users to collect and store text clips from online sources, usually from news sites, in a Personal Digital Library. The clips can then be tagged or commented by users. Our system performs an automatic text mining and clustering organization of this personal clipping collection by grouping clips with similar semantic value and taking the tags as a positive bias in the classification of content. We also provide to each user a means to dynamically change the intensity on the use of tags on the clustering process.

Unlike many other approaches [2, 7, 12], we take the view that tags may stand outside of the clustering process of the documents and form on their own an (eventually overlapping) community structure. Therefore, the discovery of this community structure of tags would be the first step in order to enhance text clustering. In this article we report an experiment where we define a distance metric based on a weighted cosine similarity, that combines the textual features with the community structure of a network of tags in order to improve the clustering of documents, in a socially biased way.

2. RELATED WORK

Using social features to improve text clustering is a subject that has recently been explored by Ares et al. [2]. They used a constrained clustering algorithm [13] to take advantage of the social tags associated with a set of bookmarked web pages in Delicious, by turning them into constraints between these documents. Simpson et al. [10] applied clustering to a tag co-occurrence graph in order to find related tags and to establish a hierarchy of tags from a flat tag list. They experimented with divisive clustering and betweenness centrality clustering, concluding that the betweenness method performed poorly on a graph of densely interconnected tags, resulting in a large dominant cluster. Han et al. [5] proposed the k-nearest neighbor classification algorithm that assigns a degree of relevance to attributes and uses them in a weighted cosine similarity measure. Yeung et al. [3] took advantage

Table 1: Average clustering coefficient and average shortest path length of the tag network and a random network with the same number of edges and nodes.

Network	Clustering Coefficient	Path Length
Random	0.04623016	3.8594669
Tags	0.85450065	3.4942602

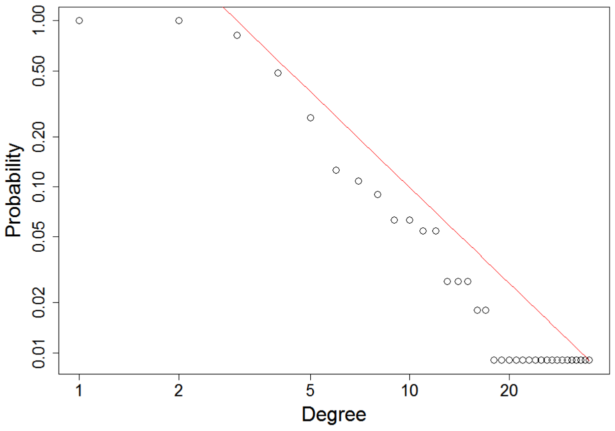

Figure 1: Degree distribution with best fit power law line.

of a folksonomy to create a disambiguation system based on the clustering of social tags. Using the clusters of the network of tags, they obtained the different groups of tags or documents associated with the distinct meanings of an ambiguous tag. Wartena et al. [14] studied the usage of tag co-occurrence for recommendation, introducing second order co-occurrence as a stable measure for tag similarities, where distances can be computed equally between users, items and tags.

We use a different approach to improve text clustering, based on the network of co-occurrence of tags. In 2008, Wu has analyzed a set of tags as a social network, revealing that it shared the traditional features of regular social networks, including the short average path length, a high clustering coefficient and a power law distribution of the node degrees [15]. They concluded that this type of networks had small world and scale-free characteristics. Based on Wu's conclusions, we apply overlapping community detection methodologies to our network of tags and use this information to determine related tags and to improve text clustering. We use the Speaker-listener Label Propagation Algorithm (SLPA) proposed by Xie et al. [16] to identify the cover (the overlapping community structure) of our network of tags.

3. A FOLKSONOMY OF NEWS CLIPS

Our data set is composed of 121 user-defined clips of news, from online sources, that were further associated with user-defined tags, having an average text length of 118.6 ± 135.5 words, and an average of 2.6 ± 0.9 tags per clip. The clips were collected in a single session by five users of the Breadcrumbs system, with a focus on five topics: the Libyan revolution, the US tax plan to tackle debt, the world debt crisis, Greek debt related events and Italy rating downgrade. Based on this data set, we construct a weighted network with tags as nodes, where each edge represents the co-occurrence count of tags in the same document. The resulting network has 111 nodes and 189 edges, a density of 0.03096 and a diameter of 19. This network displays a higher average clustering coefficient than a random network with the same node set, and low average shortest path as shown in Table 1. Also the node degree distribution follows an heavy tailed distribution as shown in Figure 1. These structural characteristics show that the graph displays community structure, making it a suitable choice for community detection.

4. NEWS CLIPS CLUSTERING

Given the community structure yielded by the structural features of the network we are able to execute an overlapping community detection algorithm over the graph to find tag communities. We use the social information yielded by the detection to improve the existing text clustering technique as part of the Breadcrumbs system. In the next sections, we describe the text clustering process and the tag clustering process, and we explain how this information is combined.

4.1 Text Clustering

The Breadcrumbs system performs text clustering of user owned collections of clips using k-means clustering with the word vectors of each documents as data points. This clustering technique classifies the collection into k different clusters of clips with related context.

The current technique makes sole use of the clip's text, which leaves out any social information that can be obtained from the clip's metadata. We integrate the folksonomy information with the current clustering technique to yield a socially improved clustering method.

4.2 Tag Clustering

Tags can have related meanings in different social contexts. Given this intrinsic property we perform overlapping community detection over the network of tags. The resulting communities represent the communities of related tags, which are sets of tags with related social context and semantics, that are used to improve the text clustering technique.

We use the overlapping community detection algorithm SLPA mainly due to its near linear complexity in sparse graphs. In principle any other algorithm capable of performing overlapping cluster/community detection should be able to perform the task in its stead. This algorithm works by propagating labels throughout each node of the network that are repeatedly stored in memory for every node. It has two parameters, a threshold r of probability, used in post-processing, and the number of iterations T. After an initialization step, SLPA starts by taking each node in a role as listener and receives one random label from each of its neighbors (speakers) and stores it in a temporary list. The listener then chooses a label from this list and adds it to its own memory according to a function based on the label occurrence count. The previous process is iterated a number of times according to the parameter, and the node memories are post-processed. The post-processing step consists of the computation of the occurrence probability for all labels and the removal from node memory of all labels with an occurrence probability below the threshold. The sets of nodes that share a certain label in its memory are constructed yielding the communities of related tags.

We introduced modifications to this algorithm to make it work with a weighted network using a modified labels list to store the sum of weights connecting to the speakers from where the label came from. The listener rule was also modified to return the label with the maximum value for the product of the sum of its weights with its occurrence count. Running the tag clustering algorithm results in nine overlapping communities, with sizes ranging from 4 to 33 nodes, where communities of 5 nodes are the most frequently identified.

4.3 Socially Biased Document Clustering

The socially biased clustering is performed by executing the modified SLPA over the network of tags with $T = 200$ iterations in the evolution step, and $r = 0.02$ for the threshold in the post-processing step. Using the community information produced, we construct the word vector for each clip according to the tf-idf score, and the tag vector for each clip using the following tag weighting function:

$$w(t,d) = (1 - SS) \times tfidf(t,d) + SS \times \frac{1}{|C_t|} \sum_{tr \in C_t} tfidf(tr,d)$$

where SS is the Social Slider, a real number between 0 and 1, with 0 disabling any biasing and 1 discarding all information but that contained in the tag clusters. This value is used to impart the tag weight function and the subsequent clustering with a quantitative social bias. C_t is the union of all overlapping communities of tags related with a tag t.

We construct the k-means data point vectors by concatenating the two vectors with a unique ID and each data point is pre-processed according to the equation:

$$v'_i = \begin{cases} v_i \times (1 - SS) & i < j \\ v_i \times SS & i \geq j \end{cases}$$

where j is the vector index of the first tag component, v the original data point, and v' the resulting data point. We then use the pre-processed data points to run k-means using a distance given by the cosine similarity proximity measure:

$$Cos_{dist}(a,b) = 1 - \left(\frac{\sum_{i=1}^{n}(a_i \times b_i)}{\sqrt{\sum_{i=1}^{n}(a_i)^2} \times \sqrt{\sum_{i=1}^{n}(b_i)^2}} \right)$$

where n is the length of the vectors a and b.

The complexity added over k-means is for the SLPA step $O(K\,|T|)$ where K is the number of iterations and T the set of tags. For the tag vector calculation, the complexity is $O(|C|\,|T|\,I)$, where C is the set of clips and I the implementation-dependent complexity of tf-idf. The preprocessing step has a complexity of $O(|C|\,|S|)$ where $|S|$ is the size of each data point.

5. EVALUATION OF CLUSTERING

We manually annotate the news clips collection, classifying each clip into one of the following six classes: Libya, US Tax, World Debt Crisis, Italy Downgrading, Greece, and Other. We use this clustering partition as our "ground truth", to which we compare the partitions resulting from the text clustering and from the combination of the text clustering with the tag clustering. In Table 2 we present the confusion matrix analysis for each of the methods, where "class" refers to our manual annotation of the clips and "cluster" refers to the partitions identified by the tested methods — Text clustering in Table 2a and Text+Tags clustering in Table 2b. The true positive rate (TPR) for the text-based method is 32.15% and the false positive rate (FPR) is 26.32%. Even though the TPR for the combined text and tags method takes a lower value of 29.70%, the FPR also decreases to 23.55%, which means that the text-based method achieves a higher number of correctly classified documents, but also a higher number of incorrectly classified documents. Since these metrics do not provide the grounds for a conclusion, we use the Rand index [9] to measure the similarity of the resulting partitions with the ground truth, i.e. the percentage of correct decisions, and the F-score to calculate the accuracy of the two methods, first using $\beta = 1$ and then using $\beta = 0.5$ and $\beta = 2$ to penalize the false negatives less and more strongly, respectively, than the false positives. Table 3 depicts the evaluation of the identified partitions using a null weight (Text), as well as a 50% weight (Text+Tags) for the social aspect. That is, for the Text clustering, we set the social slider to zero ($SS = 0$), while for the Text+Tags clustering we set the social slider to 0.5 ($SS = 0.5$), in our weighted cosine similarity proximity measure.

As we can see in Table 3b, we obtain a higher Rand index when using the social structure of the network of tags in the clustering process. On the other hand, by looking at the F-score for either method in Table 3a, we verify that using the community structure of the co-occurrence of tags in news clips slightly decreases the accuracy of the clustering method, except when given a higher weight to the precision ($\beta < 1$). Since the F-score values for the two clustering methods are very close together and the Rand index isn't by itself conclusive, we further investigate by calculating the adjusted Rand index according to Hubert & Arabie [6] and Morey & Agresti [8]. These adjusted for chance metrics are depicted in Table 3b. The resulting values are in agreement with the previously calculated Rand index, indicating that the higher Rand index for the Text+Tags clustering represented in fact a solid and significant result.

6. CONCLUSIONS

We have proposed a weighted cosine similarity proximity measure that takes into account the social information present in the underlying network of tags in a folksonomy. We used this metric as the distance function of the k-means algorithm, in order to cluster news clips together. The proposed clustering method is based on the usage of what we call a social slider, where the user can set the degree to which the social aspect of the news clipping process influences the grouping of news clips in his/her own Personal Digital Library, or across the whole system. The data points we use not only include information about the textual content and the tags of the news clips, but also about the related tags,

Table 2: Confusion matrix for the resulting partitions.

	Same cluster	Different clusters	Total
Same class	TP = 460	FN = 971	1431
Different classes	FP = 1534	TN = 4295	5829
Total	1994	5266	7260

(a) Text clustering.

	Same cluster	Different clusters	Total
Same class	TP = 425	FN = 1006	1431
Different classes	FP = 1373	TN = 4456	5829
Total	1798	5462	7260

(b) Text+Tags clustering.

Table 3: Evaluation of the clustering methods.

Clustering	$F_{0.5}$	F_1	F_2
Text	0.2444988	**0.2686131**	**0.2980047**
Text+Tags	**0.246434**	0.2632394	0.2825047

(a) F-score for $\beta = 0.5$, $\beta = 1$ and $\beta = 2$.

Clustering	Rand index	HA ARI	MA ARI
Text	0.65495868	0.05075362	0.07524322
Text+Tags	**0.67231405**	**0.05602792**	**0.08241934**

(b) Rand index and adjusted Rand indices according to Hubert & Arabie and Morey & Agresti.

which are identified based on the overlapping community structure of the global network of tags.

We performed an evaluation of the identified partitions, comparing them to our manually annotated partition. The socially biased document clustering method that we've introduced here was able to produce an improved clustering partition, by taking advantage of the social features in our documents. Identifying the overlapping community structure of the network of tags associated with the Breadcrumbs folksonomy seems to improve regular text clustering, resulting in a better grouping division of our news clips collection. We hypothesize that, as the network of tags grows and its community structure becomes stronger, groups of tags will become more cohesive and continuously result in improved socially biased clusters.

7. FUTURE WORK

As future work, we would like to replicate this experiment at a larger scale, after the data set of news clips has been increased. We believe this would reflect on the improvement of the social structure in the network of tags and therefore result in better clusters for higher values of the social slider. Additionally, we would like to test the complexity of our methodology and work on the problems that a large-scale environment introduces.

8. REFERENCES

[1] Álvaro Figueira et al. Breadcrumbs: A social network based on the relations established by collections of fragments taken from online news. *Retrieved January 19, 2012, from http://breadcrumbs.up.pt.*

[2] M. Ares, J. Parapar, and A. Barreiro. Improving Text Clustering with Social Tagging. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM 2011)*, pages 430–433, Barcelona, Spain, 2011.

[3] C. M. Au Yeung, N. Gibbins, and N. Shadbolt. Contextualising tags in collaborative tagging systems. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, pages 251–260. ACM, 2009.

[4] S. Golder and B. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[5] E. Han, G. Karypis, and V. Kumar. Text categorization using weight adjusted k-nearest neighbor classification. *Advances in Knowledge Discovery and Data Mining*, pages 53–65, 2001.

[6] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[7] X. Ji, W. Xu, and S. Zhu. Document clustering with prior knowledge. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 405–412, 2006.

[8] L. Morey and A. Agresti. The measurement of classification agreement: an adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurement*, 44(1):33–37, 1984.

[9] W. M. Rand. Objective Criteria for the Evaluation of Methods Clustering. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[10] E. Simpson. Clustering tags in enterprise and web folksonomies. *HP Labs Techincal Reports*, 2008.

[11] F. Suchanek, M. Vojnovic, and D. Gunawardena. Social tags: meaning and suggestions. In *Proceeding of the 17th ACM conference on Information and knowledge management*, pages 223–232. ACM, 2008.

[12] J. Tang. Improved K-means Clustering Algorithm Based on User Tag. *Journal of Convergence Information Technology*, 5(10):124–130, 2010.

[13] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 577–584, 2001.

[14] C. Wartena, R. Brussee, and M. Wibbels. Using tag co-occurrence for recommendation. In *Ninth International Conference on Intelligent Systems Design and Applications (ISDA 2009)*, pages 273–278. IEEE, 2009.

[15] C. Wu. Analysis of Tags as a Social Network. In *International Conference on Computer Science and Software Engineering (ICCSSE 2008)*, volume 4, pages 651–654. IEEE, 2008.

[16] J. Xie, B. Szymanski, and X. Liu. SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. *Arxiv preprint arXiv:1109.5720*, 2011.