ANNOTATION-AWARE WEB CLUSTERING BASED ON TOPIC MODEL AND RANDOM WALKS

Jiashen Sun, Xiaojie Wang, Caixia Yuan, Guannan Fang

Center of Information Science and Technology, Department of Computer Science,
Beijing University of Posts and Telecommunications, Beijing
{Jiashen_Sun, xjwang, yuancx}@bupt.edu.cn, gnfang@gmail.com

Abstract

Web page clustering based on semantic or topic promises improved search and browsing on the web. Intuitively, tags from social bookmarking websites such as del.icio.us can be used as a complementary source to document thus improving clustering of web pages. In this paper, we present a novel model which employs topic model to associate annotated document with a distribution of topics, and then constructs a graph including tags, document and topics by performing a Random Walks for clustering. We examine the performance of our model on a real-world data set, illustrating that our model provides improved clustering performance than algorithm utilizing page text alone.

Keywords: social tagging; topic model; random walks; web clustering

1 Introduction

Search engines are an invaluable tool for retrieving information from the Web. In response to an often ambiguous user query, they return a list of results ranked in order of relevance to the query. The user starts at the top of the list and follows it down examining one result at a time, until the sought information has been found. For example, when a user submits query "jaguar" into Google, s/he would want to get search results related to "big cats", "Mac OS", "car club", "sport car" etc. Recently, ranking techniques harnessing link, anchor text, and user click-through data as well as simple page text have been developed to address challenge especially in finding the home page [1]. However, these methods might be less effective for satisfying broad or ambiguous queries. One of the most promising (and common) approaches to handle this ambiguity both in query and search result is through automatic clustering of web pages in advance [5, 11, 15, 16]. This method group results to various topical clusters which can meet different information needs. Therefore, users can more quickly and accurately locate the search results representing their information need, therefore improving user experience in information retrieval.

Social annotation is a form of folksonomy, which refers to Internet-based methods for collaboratively generating short free-form text labels that categorize content such as webpage, image, or video. Many popular Web services rely on folksonomies such as del.icio.us, Digg, StumbleUpon and Flickr, collecting hundreds of thousands of keyword annotations per day [2]. The set of tags applied to a document is an explicit set of keywords that users have found appropriate for categorizing that document within their own filing system. Thus tags provide a uniquely well suited and extra source for the documents, which can be exploited in various document computing. An increasing number of studies use social annotations to optimize web search [2, 3] and clustering [5].

Incorporating social annotations with document content is a natural idea, especially for IR applications. Our goal is to investigate how annotation can best be used to improve performance of webpage clustering. We implement topic model and random walk over a <doc, tag, topic> graph in order to mine the global correlation of the documents, tags and topics. According to the steady-state probabilities, it is easy to calculate the global relevance probabilities among documents, tags or topics. Then we group the documents based on probabilities of transitioning from the document to any of nodes.

The remainder of this paper is structured as follows. We first present an overview of related work in Section 2, then describe our model in more details in Sections 3 and 4. Section 5 describes the data set and evaluation results of our experiments. We conclude our work in Section 6.

2 Related Work

To handle ambiguity both in query and search result and improve the quality of search, search result clustering is developing rapidly and much related work has been done on it in recent years [6, 7, 9]. These methods group the results returned by a search engine into a hierarchy of (or flat) labeled clusters, in that the user may focus on a general topic by a weakly-specified query, and then drill down through the highly-specific themes that have been dynamically created from the query results. However, these ways are almost non-real time and lack of unanimous gold-standard, since clustering results varies different queries.

Another approach to handle this trouble is through automatic clustering of web pages in advance. Traditional clustering algorithms [15] do not make use of the characteristics of the Web, such as hyperlink structures, anchor text and social tags. Chau et al. [11] propose to incorporate hyperlink analysis into the traditional vector space model used in document clustering. Specifically, they introduce a new metric HFIDF based on link analysis to be used with the traditional TFIDF in similarity measure in clustering algorithms. Bohunsky and Gatterbauer [16] consider the visual structure of web pages which from DOM tree for clustering. The work of Ramage et al. [5] is closest to ours. It also uses social tags as a complementary data source to page text and anchor text for improving automatic clustering of web pages. He presents Multi-Multinomial LDA that explicitly models text and tags, significantly outperforming K-means. But he fails to take into account the global correlation of documents, tags and topics.

The study of the relevance or similarity score refined by random walk has drown considerable research due to its good properties: compared with those pair-wise metrics, it can capture the global structure of the graph [4]; compared with those traditional graph distances (such as shortest path, maximum flow etc.), it can capture the multi-facet relationship between two nodes [17]. In our work, we apply random walk to cluster documents model to get a deterministic clustering on the weighted graph associated with documents, tags and topics.

3 Graph Construction

Based on the webpages with social tags and the cluster of documents, we can construct a weighted link graph G = (V, E, W) where V=T ∪ D ∪ C, represents a set of three types of object: tags, web pages, and topics of pages, as illustrated in Figure 1. Each edge in E connects the vertices in V, with W denoting the weight of edges. When constructing the link graph, tag-tag relevance and tag-topic relevance are not taken into account.

3.1 Relevance of tag and document

The most straightforward way is to directly take p(t|d) as the relevance score, since it indicates the probability of tag t given web document d. However, the tag may not be so descriptive when it appears too frequently in the dataset. For example, for the tag “Wikipedia”, the probability p(t|d) for a specific document d will be almost 1, but obviously this tag is non-informative. Therefore, we normalize p(t|d) by p(t)(Eq. 1), i.e., the prior probability of the tag, to penalize frequently-appearing tags [12].

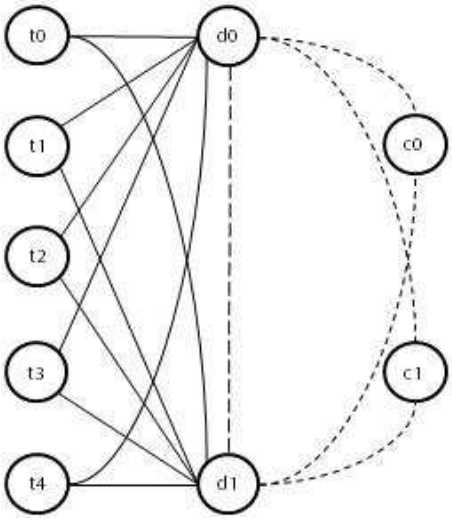

Figure 1. Annotation-Topic Graph (d, t and c corresponds to doc, tag, and topic nodes, respectively)

$$s(t,d)=\frac{p(t|d)}{p(t)} \qquad (1)$$

Based on Bayes’ rule, we can easily derive that

$$\frac{p(d|t)\,p(t)}{p(d)\,p(t)}=\frac{p(d|t)}{p(d)}\propto p(d|t) \qquad (2)$$

where p(d) is prior probability density function and p(d|t) probability density function of documents conditioned on tag t. Since the target is to rank the tags for the individual document and p(d) is identical for these tags, the relevance score is proportional to p(d|t).

We adopt the Kernel Density Estimation (KDE) method [12] to estimate the probability density function p(d|t).

$$p(d|t_i)=\frac{\sum_{d_k\in D_i}\mathrm{sim}(d,d_k)}{|D_i|} \qquad (3)$$

where $|D_i|$ is size of document set and sim() is some kind of similarity function between documents, which is mentioned in Section 3.2.3. The relevance score actually has a very intuitive explanation. A tag is representative if the entire documents annotated with the tag are similar to each other.

3.2 Document Similarities

According to Ramage et al. [5], a document with tags can be represented as a vector V in which words (in document) and/or tags are weighted by tf-idf. Since topic models have the potential to better model the data, document also can be viewed as a probability distribution over a set of latent and essential topics, which has been proved to be effective in information retrieval and text clustering [21, 22]. We here utilize a typical topic model named Latent Dirichlet Allocation (LDA) [13] to get the hidden topics of a document.

LDA models each document using a mixture over K topics, which are in turn characterized as distributions over words. We estimate the document-topic distribution $p(z_i\,|\,d)$ from an unlabeled corpus of documents using Gibbs sampling [20].

Then we use the topic proportions to define a topic-based similarity measure between documents. The vector space in topic level provides a low dimensional and simplified representation for document. One can use the Jensen-Shannon divergence (JSD, symmetric measure based on KL-divergence), Hellinger distance or cosine metrics as a similarity (or divergence) measure between two documents d_1 and d_2.

$$\mathrm{JSD}(d_1, d_2) = \frac{1}{2}\mathrm{KLD}(d_1 \,\|\, d_M) + \frac{1}{2}\mathrm{KLD}(d_1 \,\|\, d_M) \quad (4)$$

$$\mathrm{Helg\_dist}(d_1, d_2) = \sum_{k=1}^{K} (\sqrt{d_{1,k}} - \sqrt{d_{2,k}})^2 \quad (5)$$

$$\mathrm{sim}(d_1, d_2) = e^{\left(-\frac{\mathrm{dist}(d_1, d_2)}{\sigma^2}\right)} \quad (6)$$

$$\cos(d_1, d_2) = \frac{d_1.d_2}{|d_1||d_2|} \quad (7)$$

where d_1 and d_2 are the posterior topic proportions of document i.e. $p(z_i\,|\,d)$, $d_M = \frac{1}{2}(d_1 + d_2)$, KLD() is Kullback–Leibler divergence and dist() is JSD or Hellinger distance.

Similarities below a threshold (0.05 in our method) are discarded, since they would not have a notable impact on global correlation between documents.

3.3 Relevance of Topic and Document

The posterior topic proportions $p(z_i\,|\,d)$ provides a natural way to represent the relevance of document and corresponding topic.

3.4 Graph Normalization

Since the whole graph contains three types of relevance, it is necessory to normalize each relevance measures respectively.

We define a transition matrix P whose element p_{ij} indicates the probability of the transition from vertex i to vertex j, computed as

$$p_{ij} = \frac{s_{ij}}{\sum_k s_{ik}} \quad (8)$$

where s_{ij} denotes the relevance score between node i and j. The transition probability is calculated by normalizing the score out of node i.

Let W represent an (|D|+|T|+|Z|)×(|D|+|T|+|Z|) weight matrix, in which element is transition probability between any two nodes.

$$W = \begin{pmatrix} P_{DD} & P_{DT} & P_{DZ} \\ P_{TD} & P_{TT} & P_{TZ} \\ P_{ZD} & P_{ZT} & P_{ZZ} \end{pmatrix} \quad (9)$$

There are several ways to normalize the weighted matrix W. The most natural way might be by row normalization [18, 19]. Complementarily, Tong et al. [17] propose using the normalized graph Lapalician ($W=D^{-1/2}WD^{-1/2}$) as an alternative. In our experiment, Lapacian normalization shows better performance than row normalization.

4 Random Walk with Restart

Random walk (RW) methods have been widely applied in machine learning and information retrieval fields [4, 8, 17]. In order to mine the global relationship of the documents, tags and clusters, we perform random walk over the <tag, doc, cluster> graph constructed in Section 3, and then group documents based on the stationary distribution.

4.1 Computing the Stationary Distribution

Consider a random particle that starts from node i. The particle iteratively transmits to its neighborhood with the probability that is proportional to their edge weights [17]. Also at each step, it has some probability c to return to the node i (named restart probability). The relevance score of node k to node i is defined as the steady-state probability p_{ki} (see Eq. 10) that the particle will finally stay at node i.

$$r_i = \left(p_{1i}, \ldots,\ p_{ki}, \ldots\ p_{|V|i}, \right)^T \quad (10)$$

$$r_i = cWr_i + \left(1 - c\right)e_i \quad (11)$$

Eq. 11 defines a linear system problem, where r_i is determined by:

$$r_i = \left(1 - c\right)(I - cW)^{-1}e_i = \left(1 - c\right)Q^{-1}e_i \quad (12)$$

It can be observed from Eq. 12 that the system matrix Q defines all the steady-state probabilities of random walk with restart. Thus, if we can pre-compute and store Q^{-1}, we can get r_i real-time.

4.2 Clustering

According to the steady-state probabilities, it is easy to calculate the probability of transitioning from a certain document to any other node. We examine two strategies to obtain clustering result, a) generate the most probable topic as result for the document directly, b) group document based on structural similarity again using a second clustering algorithm.

5 Experiments

5.1 Gold Standard

We derive gold standard clusters from the Open Directory Project (ODP). All top-level categories, their children and children's descendants were regarded as the gold standard of category group, except Regional and World category, because they are categorized primarily by region, not by topic. URLs poitering to homepage and multimedia page are filtered out. Finally, we have about 420,000 URLs.

To get the social annotation of these pages, we obtain 110,000 URLs which are included in ODP and also tagged in del.icio.us. After HTML parsing, only 60,000 were in English and had their page text crawled by wget. Through an indispensable tag pre-filtering process, 8,328 document with social tags and category are gathered as gold standard data. The category distribution of documents is shown in Table 1. We can observe that documents are seriously unbalanced over categories which highlights the great challenges to clustering task.

Category	#docs
Arts	1692
Business	164
Computers	935
Games	92
Health	427
Home	344
News	3
Recreation	11
Reference	13
Science	2328
Shopping	21
Society	2095
Sports	203

Table 1 Category distribution of Experimental data

5.2 Evaluation Metric

V-measure [14] is an entropy-based measure which explicitly measures how successfully the criteria of homogeneity and completeness have been satisfied. Homogeneity assesses the degree to which each cluster contains instances from a single class of C. Completeness assesses the degree to which each class is contained in a single cluster. As F-measure scores can be weighted, V-measure can be weighted to favour the contributions of homogeneity or completeness.

V-measure has important advantages over popularly used F-measure [10]: it does not assume a mapping between classes and clusters and their scores depend only on the relative sizes of the clusters. In this work, we evaluate our results using V-measure described above.

$$h = 1 - \frac{H(C|K)}{H(C)}$$
$$c = 1 - \frac{H(K|C)}{H(K)} \qquad (16)$$
$$V_\beta = \frac{(1+\beta)*h*c}{(\beta*h)+c}$$

5.3 Baseline

In our experiments, we look at two broad families of baseline clustering algorithms. The first family is a k-way N-cut graph partitioning, which has the best performance in our baseline experiments. The documents are based on word-level VSM and topic distribution mentioned in Section 3.2. We examine two ways to model a document based on BOW [5]:

Words Only: V_w is defined as $(w_1, w_2, \ldots, w_{|W|})$ where w_j is the tf-idf assigned to word j.

Tags + Words: If we define V_w to be the words only vector and V_t to be the tags only vector, then the Tags + Words vector $V_{w+t} = \langle V_w, V_t \rangle$. In other words, the vector is modeled as concatenation of word and tag, giving words and tags equal weight.

The second family is based on the clustering property of topic model. We are more interested in estimating p(z|d), the topic-document distribution mentioned in Section 3.3. Our system labels each document with the single, most probable topic as the clustering result directly, named Direct_LDA.

5.4 Experiments Results

Our result is averaged results of 5-fold cross validation. The topic number in LDA is set 20. Table 2 demonstrates the performance of our experiments in different similarity (or divergence) measure.

	V-measure
LDA (cos)	28.2
LDA (Hellg)	30.5
LDA (JSD)	30.0

Table 2 V-measure of different similarity (or divergence) measure

As shown in Table 2, Hellinger distance (Hellg) and JSD are more effective than cosine, while Hellinger distance is the best suited to measure similarity (or divergence) between two probability distributions in our experiments.

Table 3 show that K-way N-cut on Words+Tags outperforms that on words alone, indicating that tags are a qualitatively different type of content which can effectively improve web clustering. In addition, K-way N-cut on topic-level similarity significantly improves the performance as we have expected. The most likely reason is that topics can effectively exploit inherent information of document. It is worthwhile to note that Direct_LDA also achieves a satisfactory V-measure when compared with K-way N-cut model.

Finally, it is found that our LDA_RW_Ncut method significantly outperforms all other configurations, which meet our initial assumptions.

	V-measure
Words Only	23.3
Words+Tags	24.6
LDA (Hellg)	30.5
Direct_LDA	30.3
Direct_LDA_RW	32.4
LDA_RW_Ncut	36.8

Table 3 V-measure of different clustering strategies, where Direct_LDA means "most probable topic as the clustering result directly", Direct_LDA_RW means "most probable topic (after Random Walk) as the clustering result directly" and LDA_RW_Ncut means "group document based on structural similarity again using Ncut".

6 Conclusions

This paper investigated the problem of web text clustering with annotation. We had shown that social tagging data provides a useful source of information for web page clustering. We thus proposed an annotation-aware clustering approach to makes even better use of global and smoothed similarity information held in tags, words and topics of a document. Experiment results demonstrated that our novel approach significantly outperforms algorithm utilizing page text alone.

In future work, we will examine other topic models such as MM-LDA or MM-HDP in order to better exploit the inherent relationship among webpages, annotations and topics. Meanwhile, it is needed to study more extensively that whether some other features like keyword in webpages are benefit to improve web clustering.

Acknowledgements

This work is supported by NSFC90920006 and RFDP 20090005110005.

References

[1] C. D. Manning, P. Raghavan, H. Schütze. Introduction to Information Retrieval. Cambridge University Press. 2008

[2] P. Heymann, G. Koutrika, H. Garcia-Molina. Can Social Bookmarking Improve Web Search. WSDM 2008.

[3] Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, and Zhong Su. Optimizing web search using social annotations. WWW. 2007.

[4] J. He, M. Li, H. Zhang, H. Tong, and C. Zhang. Manifold-ranking based image retrieval. ACM Multimedia. 2004.

[5] D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina. Clustering the tagged web. WSDM. 2009

[6] H.J. Zeng, Q.C. He, Z. Chen, W.Y. Ma, and J. Ma. Learning to cluster web search results. SIGIR, pages 210-217, 2004.

[7] O. Zamir and O. Etzioni. Web document clustering: A feasibility demonstration. SIGIR. 1998.

[8] D. Ramage, A. N. Rafferty, and C. D. Manning. Random Walks for Text Semantic Similarity. ACL-IJCNLP TextGraphs-4 Workshop. 2009.

[9] X. Wang and C. Zhai. Learn from Web search logs to organize search results. SIGIR. 2007.

[10] C. J. Van Rijsbergen. Information Retrieval, 2nd edition. 1979.

[11] Chau, M., Chau, P. Y. K., and Hu, P. Incorporating Hyperlink Analysis in Web Page Clustering. Proceedings of the Sixth Workshop on E-Business (WEB 2007). 2007.

[12] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag Ranking. Proc. WWW'09. 2009.

[13] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research. 2003.

[14] A. Rosenberg and J. Hirschberg. Vmeasure: A conditional entropy-based external cluster evaluation measure. Proceedings of EMNLP-CoNLL. 2007.

[15] P. Ruhlen, H. T. Ilhan, V. Livshits. Unsupervised Web Page Clustering. 2000.

[16] W. Gatterbauer, P. Bohunsky. Visual Structure-based Web Page Clustering and Retrieval. WWW poster track. 2010

[17] Hanghang Tong, C. Faloutsos, Jia-Yu Pan. Fast Random Walk with Restart and Its Applications. ICDM. 2006

[18] T. Hughes and D. Ramage. Lexical Semantic Relatedness with Random Walks. EMNLP. 2007

[19] Tae Hoon Kim, Kyoung Mu Lee, and Sang Uk Lee. Generative Image Segmentation Using Random Walks with Restart. ECCV. 2008

[20] T. L. Griffiths and M. Steyvers. Finding scientiffic topics. Proceedings of the National Academy of Sciences. 2004.

[21] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. WWW. 2008.

[22] X. Wei and W.B. Croft. LDA-based document models for ad-hoc retrieval. *SIGIR '06*. 2006