Topical Key Concept Extraction from Folksonomy

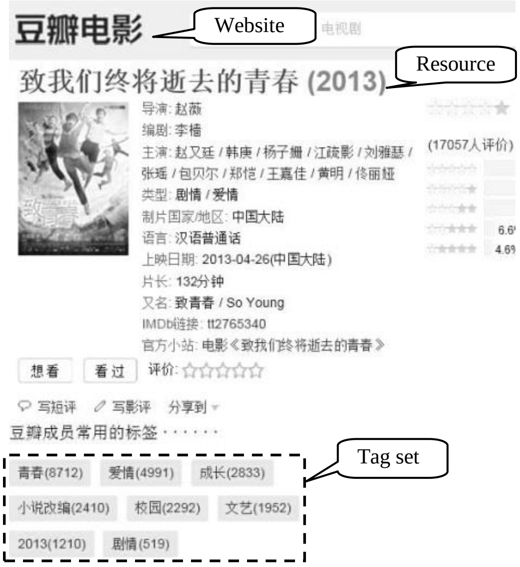
Han Xue[1,2], **Bing Qin**[1*], **Ting Liu**[1], **Chao Xiang**[1]
[1]Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, Harbin, China
{hxue,bqin,tliu}@ir.hit.edu.cn, Cloudaice@gmail.com
[2]Harbin Engineering University, Harbin, China

Abstract

Concept extraction is a primary subtask of ontology construction. It is difficult to extract new concepts from traditional text corpus. Moreover, building a single ontology for multiple-topic corpus may lead to misconception. To deal with these problems, this paper proposes a novel framework to extract topical key concepts from folksonomy. Folksonomy is a valuable data source due to real-time update and rich user-generated contents. We first identify topics from folksonomy using topic models. Next the tags are ranked according to their importance for a certain topic by applying topic-specific random walk methods. The top-ranking tags are extracted as topical key concepts. Especially, a novel link weight function which combines the local structure information and global semantic similarity is proposed in importance score propagation. From the perspectives of qualitative and quantitative investigation, our method is feasible and effective.

1 Introduction

Ontology can be seen as an organized structure of concepts according to their relations (Cui et al., 2009). Therefore, concept extraction is an important subtask of ontology construction. Existing works mainly focus on extracting concepts from text corpus (Buitelaar et al., 2005). However, it is difficult to find text corpus that accurately characterize a highly focused, even fast-changing topic (Liu et al., 2012) of the domain. For instance, it is easier to find text corpus for a common topic of movie such as "comedy", but it is much more difficult to find one for a specific topic such as "cult". Since "cult" movies often do not follow traditional standards of mainstream movies. Moreover, it is not easy for them to find a formal definition and description in text corpus. However, we can more easily find some tags (arbitrary words assigned by people to the resources of interest) to describe this kind of movie, such as cult, non-mainstream, small budgets and so on. Motivated by the fact that social tags give us flexibility and ease to describe a topic, we try to use folksonomy (Trant, 2009) as a new data source. The word 'folksonomy' is a blend of the words 'folk' and 'taxonomy'. It is the achievement of collective wisdom derived from the practice of collaboratively creating tags to annotate and categorize web resource.



致我们终将逝去的青春 (2013)

导演: 赵薇
编剧: 李樯
主演: 赵又廷 / 韩庚 / 杨子姗 / 江疏影 / 刘雅瑟 / 张瑶 / 包贝尔 / 郑恺 / 王嘉佳 / 黄明 / 佟丽娅
类型: 剧情 / 爱情
制片国家/地区: 中国大陆
语言: 汉语普通话
上映日期: 2013-04-26(中国大陆)
片长: 132分钟
又名: 致青春 / So Young
IMDb链接: tt2765340
官方小站: 电影《致我们终将逝去的青春》

豆瓣成员常用的标签 · · · · · ·

青春(8712) 爱情(4991) 成长(2833)
小说改编(2410) 校园(2292) 文艺(1952)
2013(1210) 剧情(519)

Figure 1. A folksonomy example

Take Douban.com for example, which is a Chinese SNS website allowing registered users to record information and create tags related to

[*]Correspondence author

their interested resources, such as film, books, music and recent activities. As shown in Fig. 1, in the folksonomy-driven web site 豆瓣电影网站‘Douban.com Movie[1]’, the resource 致我们终将逝去的青春‘So Young’ is annotated with a set of tags including 青春‘youth’, 爱情‘romance’, and 成长‘growth’ ordered by the frequency of use which update automatically.

Compared with traditional text corpus, folksonomy can overcome the knowledge acquisition bottleneck. It is superior to text corpus in three aspects. (1) tag is more free and easier to characterize a highly focused, even fast-changing topic; (2) tag as a candidate concept has been extracted by collective wisdom, which avoids a series of natural language processing tasks applied to text corpus such as word segmentation, part of speech tagging, and syntactic parsing and so on; (3) the associated relationships among resources, tags and users through tagging provide a large amount of potentially valuable semantic information for mining. However, folksonomy also has two disadvantages, such as ambiguity and lack of hierarchy. To avoid misconception, we think of building multiple topic-specific ontologies instead of a single one.

In this paper, we propose to automatically extract topical key concepts from folksonomy. The topical key concept should be abstract, representative of the corresponding topic. It should contain common features that can be inherited by other non-core relevant concepts under the same topic. For example, the topical key concepts in the field of movie may be comedy, biography and action and so on. To extract the topical key concepts, we learn the topic distribution of the tags by applying LDA (Latent Dirichlet Allocation) (Blei et al., 2003) at first. After that, the tags are ranked on the basis of the importance scores for a certain topic by a variant of topic-specific PageRank (Page et al., 1999). Specially, the novel contribution of the variant is a new link weight function in importance propagation, which combines the local similarity (defined as co-occurrence of tags in a same resource assigned to the given topic) with the global similarity (defined as cosine similarity of two tags over all the topic dimensions in the whole collection considered). Then, the top-ranking tags that best represent the corresponding topic are extracted as topical key concepts.

In view of limited Chinese corpus and complex Chinese syntax for ontology construction, we tried on Chinese folksonomy data. Experiments on movie data from Douban.com show that new link weight function can largely help boost the performance. To the best of our knowledge, our work is the first to study how to extract topical key concepts from folksonomy in the field of Chinese ontology construction. We perform a thorough analysis of the proposed method, which can be useful for future work in this direction.

Although our goal is to build Chinese ontology based on the topical key concepts from this work, our method can be widely used in many other tasks such as information navigation and recommendation system. Furthermore, our method is unsupervised and language independent, which is applicable in the web era with enormous information.

The rest of the paper is organized as follows. Section 2 reviews some related works; Section 3 describes our proposed method; Section 4 presents our experiments and resultant analysis; and Section 5 draws the conclusions and directions for the future work.

2 Related Work

Many efforts have been made to extract the key concepts for ontology construction. These methods can be divided into two categories according to topic-sensitive or not.

Topic-free Some key concepts of famous ontologies are usually defined by linguists or domain experts. The suggested upper merged ontology (SUMO) is such a kind of ontologies. The expert-based methods are accurate and standard. However, to tackle the time-consuming and laborious problems, efforts are also made to use semi-automatic and automatic methods.

Among semi-automatic methods, rule-based methods are known for high accuracy if the patterns are carefully chosen according to morphological structure or special format of corpus (Nakayama et al., 2008), either manually or via automatic bootstrapping (Hearst, 1992). However, the methods suffer from sparse coverage of patterns in a given corpus.

Some researchers try to map the words to a thesaurus or an existed ontology (WordNet or Wikipedia) automatically so as to get key concepts (Angeletou et al., 2008). The coverage and openness of existed ontologies seriously limit the scope of these works. Simple statistical methods

[1] http://movie.douban.com

such as TF-IDF weighting (Hulth, 2003) are not feasible for folksonomy since short text snippets only. Graph-based ranking methods are the state of the art. They are superior to the statistic-based methods because of considering structure information between words. Mihalcea and Tara (2004) propose to use TextRank, a modified PageRank algorithm to extract key concepts from text. But TextRank only maintain a single importance score for each word. Hotho et al. (2006) propose a graph-based ranking algorithm for folksonomy, named FolkRank. They convert triadic hypergraph in folksonomy into an undirected tripartite graph. But we consider that the tripartite graph may include much noise for key concept extraction.

As a word usually spans multiple topics, the importance of the word with respect to different topics would be different. It seems that the previous works mentioned above may lead to misconception by mixing different topics together.

Topic-sensitive In order to overcome misconceptions, the topic models and other clustering models such as DA (Deterministic annealing) (Zhou et al., 2007) are used to derive topical key concepts from corpus based on word occurrence information. These clustering models usually regard corpus as a bag of words. They can find the topic or the leading word in each cluster, but cannot distinguish concrete entity well.

It is intuitive to consider topic information in graph-based ranking methods for topical key concept extraction.

Haveliwala (2002) proposes topic-sensitive PageRank (TSPR) to get a set of PageRank vectors biased a set of representative topics and generate more accurate rankings than a single PageRank vector. Nie et al. (2006) propose a topical link analysis model (TLA) to affect the importance propagation. However, the topics in TSPR and TLA are both from ODP (Open Directory Project)[2] extracted manually. Based on their works, Jin et al. (2011) implement a topic-sensitive tag ranking (TSTR) approach in folksonomy automatically through LDA. TSTR performs better than TSPR and TLA because topics extracted by LDA are more conformed to the actual situation than the topics of ODP. They pay more attention to the effect of the transfer action probability on the importance score of tags which benefit us to know the propagation process. It seems that mixing together the random walks of all the topics in one graph may cause noise compared to independent topic graphs.

Liu et al. (2010) decompose a traditional random walk into multiple random walks specific to various topics, named Topical PageRank (TPR). The novel contribution is the study on topic-specific preference value setting. And then, Zhao et al. (2011) argue that context-free propagation may cause the importance scores to be off-topic. They model the score propagation with topic context when setting the link weights and then denote this context-sensitive topical PageRank as cTPR. Enlightened by TPR and cTPR, we further propose a new link weight function to express the semantic similarity between two tags of folksonomy. The novel link weight function combines the local similarity (defined as co-occurrence of tags in a same resource assigned to the given topic) with the global similarity (defined as cosine similarity of two tags over all the topic dimensions in the whole collection considered).

3 Method

In this section, we will introduce our method. We firstly give some definitions and then overview our method, and finally introduce topic identification and tag ranking in detail.

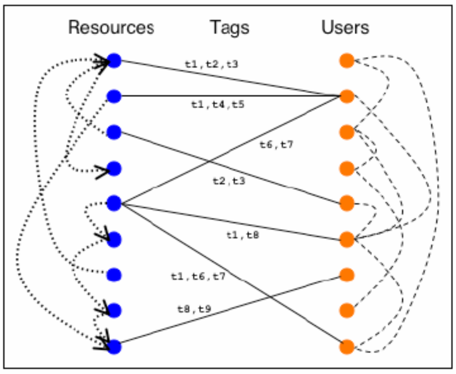


Figure 2. A conceptual model of folksonomy

From the Fig. 2 (Marlow, 2006), we can see a conceptual model of social tagging behavior in folksonomy. It consists of users $u \in U$, tags $w \in V$, and resources $s \in S$ .The conceptual model illustrates visually the implicit association among resources, tags and users joined with straight lines and dashed lines. Folksonomy is composed of < U, V, S > triples. For simplicity, we only regard a resource $s \in S$ as a document which includes a set of tags assigned by all the users of collection U. Moreover, we suppose that there is

[2] http://www.dmoz.org/

a set of topics Z over the resource collection S. Hence, we extract topics from S.

The basic idea of our method is to incorporate topic distribution into importance score propagation of tags when setting the link weight as well as the preference value. Especially, the link weight function considers both the local and global similarity with respect to different topics.

First of all, we identify topics from folksonomy resource collection S using topic models (Section 3.1). Next for each topic, we build the tag graph and use the random walk techniques to measure the tag importance. Based on the importance scores, we extract the top-ranking tags as the topical key concepts (Section 3.2).

3.1 Topic Identification

Latent Dirichlet Allocation (LDA) is a typical representative of topic models. In LDA, each word w in a document d is regarded to be generated by first sampling a topic z from d's topic distribution θ, and then sampling a word from the distribution over words Φ that characterizes topic z. θ and Φ are drawn from conjugate Dirichlet priors α and β, separately.

We use resource set S represented by a set of tag as our input file to run LDA. After the parameters converge by Gibbs sampling, we mainly use two of these output files in this paper. One is model-final.phi, which is a |Z|*|V| matrix about Φ, whose element is the probability of tag w_i conditional on topic z_j i.e. $P(w_i \mid z_j)$ .The other is model-final.tassign, which is a |S|*|V| matrix, where each row of data stands for a resource s followed by a set of elements, and each element consists of a tag and a topic which the tag most likely to be assigned to.

Through LDA, we can obtain the topic distribution of each tag $w_i \in V$ by Eq. 1, namely $P(z \mid w_i)$ for given topic $z \in Z$,

$$P(z \mid w_i) = \frac{P(z)P(w_i \mid z)}{\sum_{z'} P(z')P(w_i \mid z')} \quad (1)$$

where $P(w_i \mid z)$ can be found in the model-final.phi directly, and $P(z)$ is calculated by Eq. 2.

$$P(z) = \frac{C(z)}{\sum_{z'} C(z')} \quad (2)$$

In which, $C(z)$ is calculated as the number of times topic z appears in the model-final.tassign. Obviously, $\sum_{z'} C(z')$ is calculated as the number of times all the topics appear in the model-final.tassign. Then, we can calculate the local and global similarity between two tags using Eq. 3 and Eq. 4.

$$Local_s(w_j, w_i) = \frac{C^S_{w_j,w_i,z}}{C^S_{w_j,w_i}} \quad (3)$$

$$Global_s(w_j, w_i) = \frac{\sum_z^S p(z \mid w_j) p(z \mid w_i)}{\sqrt{\sum_z^S p(z \mid w_j)^2 \sum_z^S p(z \mid w_i)^2}} \quad (4)$$

Among them, $Local_s(w_j, w_i)$ stands for the local semantic similarity between tag w_i and w_j. In Eq. 3, $C^S_{w_j,w_i,z}$ counts the number of co-occurrences of tag w_i and w_j in a same resource of S assigned to the topic z. $C^S_{w_j,w_i}$ counts the number of co-occurrences of tag w_i and w_j in a same resource of S. We can get them from statistical calculation of the model-final.tassign. $Global_s(w_j, w_i)$ stands for the global semantic similarity between tag w_i and w_j. From the whole resource collection S, we can get the cosine similarity between tag w_i and w_j over all the topic dimensions by plugging Eq. 1 into Eq. 4.

3.2 Tag Ranking

After topic identification, we perform topical key concept extraction followed by two steps, namely, tag graph construction and tag ranking.

Above all, some formal notations are given. We denote G = (V, E) as the graph composed of tags, with vertex set $V = \{w_1, w_2, ..., w_N\}$ and link set $(w_i, w_j) \in E$ if there is a link from node w_i to w_j. In a tag graph, each vertex represents a tag, and each link indicates the correlation between every two tags. We denote the weight of the link (w_i, w_j) as $e(w_i, w_j)$, and the out-degree of vertex w_i as $O(w_i) = \sum_{j: w_i \to w_j} e(w_i, w_j)$.

PageRank assigns global importance scores to vertices using link information. In PageRank, the score $R(w_i)$ of the word w_i is defined as

$$R(w_i) = \lambda \sum_{j: w_j \to w_i} \frac{e(w_j, w_i)}{O(w_j)} R(w_j) + (1-\lambda)\frac{1}{|V|} \quad (5)$$

where damping factor λ ranges from 0 to 1, and $|V|$ is the number of vertices. The damping factor indicates that each vertex has a probability of $(1-\lambda)$ to perform a random jump to another vertex within this graph while has a probability of λ to follow the out-degree link. The PageRank importance scores are obtained by running Eq. 5 iteratively until convergence. The second term in Eq. 5 can be regarded as a smoothing factor to make the graph fulfill the property of being aperiodic and irreducible, so as to guarantee that PageRank converges to a unique stationary distribution.

In fact, the second term of PageRank in Eq. 5 can be set to be non-uniformed. The idea of Topical PageRank (TPR) is to run Biased PageRank for each topic separately. Formally, in the PageRank of a specific topic z, they assign a topic-specific preference value $\Pr_z(w)$ to each word w as its random jump probability with $\sum_{w \in V} \Pr_z(w) = 1$. For topic z, the topic-specific PageRank importance scores are defined as follows,

$$R_z(w_i) = \lambda \sum_{j:w_j \to w_i} \frac{e(w_j, w_i)}{O(w_j)} R_z(w_j) + (1-\lambda)\Pr_z(w_i) \quad (6)$$

However, TPR ignores the topical context in the link weight settings; the link weight $e(w_j, w_i)$ in Eq.6 is calculated as the number of co-occurrences of two words within a certain window size. Zhao et al. (2011) propose to use a topical context-sensitive PageRank method (cTPR). Formally, they have

$$R_z(w_i) = \lambda \sum_{j:w_j \to w_i} \frac{e_z(w_j, w_i)}{O_z(w_j)} R_z(w_j) + (1-\lambda)\Pr_z(w_i) \quad (7)$$

where they calculate the propagation from w_j to w_i in the context of the topic z, namely, the link weight $e_z(w_j, w_i)$ from w_j to w_i is parameterized by z. In Eq.7, the link weight between two words is calculated as the number of co-occurrences of the two words in tweets assigned to the topic z.

However, we believe that the link weight only contains the co-occurrence information is not enough. On the basis of cTPR, we further propose a new link weight function (Eq.8).

$$e_z(w_j, w_i) = Local_s(w_j, w_i)((1-\rho)Global_s(w_j, w_i) + \rho) \quad (8)$$

The weight factor ρ controls the proportion of the local structure information (Eq.3) and global semantic similarity (Eq.4) in Eq.8. Through the new link weight, the propagation of our method not only reflects the specific local co-occurrence information of two tags in a single resource, but also reflects the non-specific global semantic similarity of two tags in the whole resource set.

In our tag ranking method, we obtain the importance scores for each tag in different topics by Eq.7. In which, $e_z(w_j, w_i)$ is calculated by Eq.8, and $O_z(w_j) = \sum_{i:w_j \to w_i} e(w_j, w_i)$. The topic-specific preference value $\Pr_z(w_i) = P(z \mid w_i)$ is calculated as Eq.1, which is the best one among the three choices discussed by Liu et al. (2010).

4 Experiments

4.1 Dataset

Our evaluation dataset is crawled from Douban.com Movie, which is a popular Chinese Social Networking Service (SNS) website allowing registered users to create content related to movies. The dataset contains top 250 movies with 1760 tags assigned by users up to June 2012. After removing stop words and noises, we prepare 1737 tags corresponding to 249 movies for LDA. Empirically, we set the number of topics to 40 and ran LDA with 1000 iterations of Gibbs sampling.

We further select two baseline methods that most similar to ours, i.e., TPR and cTPR. All of them are iterative algorithms. We terminate the algorithms when the number of iterations reaches 100 or the difference of importance scores about each vertex between two neighbor iterations is less than 0.000001.

There are three parameters in our method that may affect the performance of the topical key concept extraction including (1) damping factor λ that reconciles the influence of adjacent nodes' importance (the first item in Eq. 7) and preference value (the second item in Eq. 7) to the modified PageRank importance of our method; (2) weight factor ρ that controls the proportion of the local structure information (Eq. 3) and global semantic similarity (Eq. 4) on two tags; (3) threshold Q; If the global semantic similarity between two tags is less than Q, we will remove the link between them. We separately set parameters λ , ρ and Q from 0.1 to 0.9 with a step size of 0.1, and then each parameter has 9 candidate values. Finally, 729 experiment results of the baseline and our method based on permutation and combination of the three parameters are presented.

4.2 Gold Standard Annotation

We construct the evaluation standard by pooling (Voorhees et al., 2005) method. The reason lies in two aspects. One is that there is no existing gold standard for topical key concept extraction from folksonomy, and the other is that it is impossible to determine all the topics and key concepts manually. We randomly mix 729 results from TPR, cTPR and our method, and then ask two judges to score as 1(relevant, abstract and representative) or 0 (irrelevant or too specific). Only if the two judges score 1 for the same tag, the tag will be determined as correct topical key concept. Otherwise, the tag will be determined as wrong.

4.3 Evaluation Metrics

The traditional evaluation metrics represented as follows,

$$P = \frac{C_{correct}}{C_{extract}},$$
$$R = \frac{C_{correct}}{C_{s\tan dard}},$$
$$F = \frac{2PR}{P+R} \quad (9)$$

where $C_{correct}$ denotes the number of correct topical key concepts extracted by a method, $C_{extract}$ denotes the number of automatically extracted topical key concepts by a method, and $C_{s\tan dard}$ denotes the total number of topical key concepts referenced by gold standard. Three of them are averaged on all the topics.

In addition to the traditional metrics precision/recall/F-measure, we use another two metrics to take the order into account.

One metric is mean average precision (MAP). MAP is desirable to measure the overall performance of topical key concept ranking,

$$MAP = \frac{1}{|Z|}\sum_{z\in Z}\frac{1}{N_z}\sum_{j=1}^{|M_z|}\frac{N_{M,z,j}}{j}I(score(M_{z,j})\geq 1) \quad (10)$$

where $I(S)$ denotes an indicator function which returns 1 when S is true and 0 otherwise, $N_{M,z,j}$ denotes the number of correct key concepts among the top j candidates returned by method M for topic z, and N_z denotes the total number of correct key concepts of topic z referenced by the gold standard.

The other metric is mean reciprocal rank (MRR) (Voorhees, 1999) which is used to evaluate how the first correct topical key concept for each topic is ranked. For a topic z, $rank_z$ is denoted as the rank of the first correct topical key concept with all extracted candidates, MRR is defined as follows,

$$MRR = \frac{1}{|Z|}\sum_{z\in Z}\frac{1}{rank_z} \quad (11)$$

where Z is the topic set for topical key concept extraction.

4.4 Quantitative Evaluation

As for the same folksonomy dataset from Douban.com Movie, we realize the baseline methods, i.e., TPR and cTPR. The TPR calculates the co-occurrence number of two tags in a same resource as the link weight, namely $e(w_j, w_i) = C^S_{w_i,w_j}$. In cTPR, the link weight is calculated as the co-occurrence number of two tags in a same resource assigned to the same given topic, namely $e_z(w_j, w_i) = C^S_{w_i,w_j,z}$.

The grid-search algorithm is applied to obtain the optimal parameter combination from 729 candidates. Through an exhaustive combination of three parameters, we obtain the best value of every evaluation indicator in three methods.

Method	P	R	F	MRR	MAP
TPR cTPR	0.617 0.625	0.404 0.406	0.465 0.473	0.670 0.675	0.405 0.407
Our method	**0.700**	**0.440**	**0.518**	**0.713**	**0.440**

Table 1. Comparisons of our method and the baselines (t-test, p-value<0.0001)

The comparison of our method to the baselines is shown in table 1. Our method achieves a 7.5% improvement in Precision over the cTPR and 8.3% over the TPR, also increased by more than 3.3% in other indicators.

We also investigate the influence of different parameter values. Due to space limitation, we only provide comparison analysis on MRR. As shown in Fig.3, the bar chart illustrates that the comparisons of our method to the baselines on MRR when λ is set 0.1, 0.3, 0.5, 0.7, 0.9, while the curve diagram describes the fluctuation of the methods. Although our method is influenced by parameter λ, ours is superior to the baselines in all parameter values.

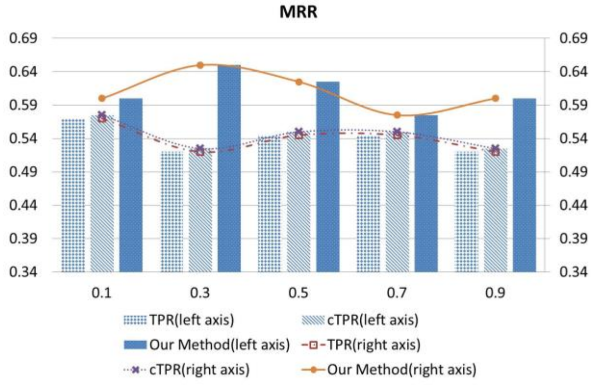


Figure 3. Comparison of the methods on MRR while varying λ (t-test, p-value<0.0001)

Similarly, we can see clearly from Fig.4 that significant promotion of our method compared to the baselines through introducing ρ. In addition, our method keeps stable with the variations of ρ.

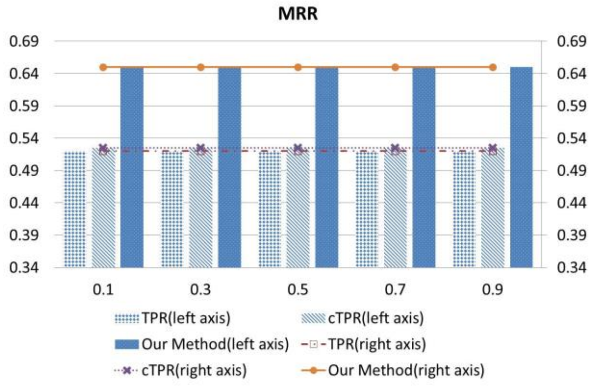


Figure 4. Comparison of the methods on MRR while varying ρ (t-test, p-value<0.0001)

The statistics in Fig.5 illustrate that our method remains stable when Q is from 0.1 to 0.7, and the curve improves significantly by increasing Q until Q reaches 0.9. While the other two methods change greatly as Q varies. Especially, the baseline methods become rather poor when Q is equal to 0.9. We infer that the baseline methods may lose many links in the tag graph when faced with a high threshold Q, because the link between two tags which the global semantic similarity lower than Q will be removed. Nevertheless, our method can deal with this problem very well. These experimental results demonstrate the robustness of our method.

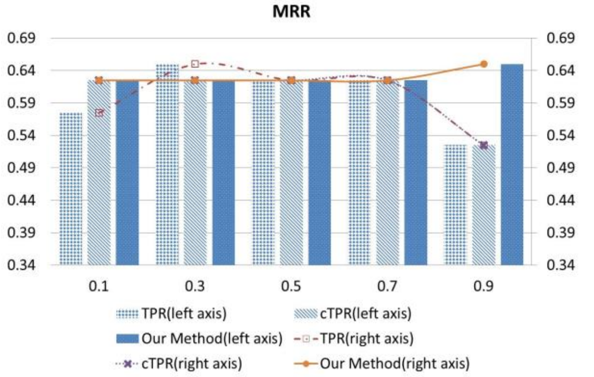


Figure 5. Comparison of the methods on MRR while varying Q (t-test, p-value<0.0001)

4.5 Qualitative Evaluation

In this subsection, some qualitative evaluations are provided on the basis of the resultant graphs with respect to different topics generated by our method.

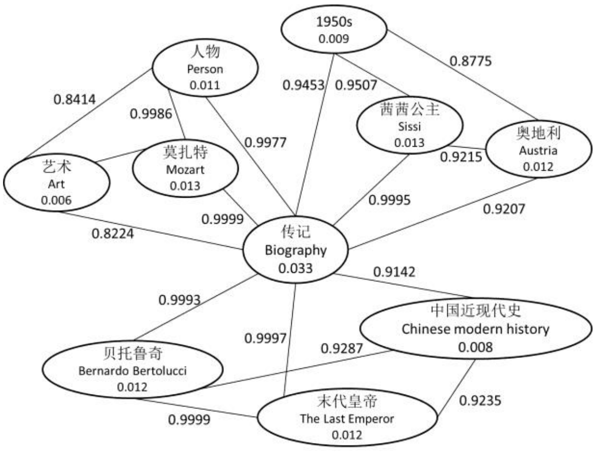


Figure 6. An example of topical graphs generated by our method

The vertex is composed of the tag name and the topical importance score, while the link weight value stands for the degree of semantic similarity between two vertices. As shown in Fig. 6 about biography topic, we can intuitively observe that the vertex 传记‘Biography’ stands in the center while surrounded by a wealth of connectivity. Compared with other vertices, it has the highest importance score in this topic and on behalf of this topic. These observations also confirm the effectiveness of our method.

To our surprise, it seems that some apparently unrelated tags are closely connected in our resultant graph. The interesting findings help us to further detect the fact that usually obtained through common sense or reasoning. For example, 末代皇帝‘The Last Emperor’ and 贝纳尔多·贝托鲁奇‘Bernardo Bertolucci’, which can be used to infer the fact that ‘The Last Emperor’ is a biographical film directed by ‘Bernardo Bertolucci’. Likewise, we observe that the dense connections among 茜茜公主‘Sissi’, 奥地利 ‘Austria’ and “1950s” can help us to infer the fact that ‘Sissi’ is a “1950s” film in ‘Austria’. All of them are connected to ‘Biography’, which means ‘The Last Emperor’ and ‘Sissi’ all belong to ‘Biography’. These insights further prove that our method can well connect the most related tags together with respect to the topic through the novel link weight model.

In addition, a few unusual genres of movie emerge in our works which enrich the traditional movie categories. For example, 公路‘road movie’, 默片‘silent movie’, 黑色电影‘film noir’, and 神片 shen-pian and so on. The sensibility for upcoming concepts indicates that our method is a necessary complement for traditional concept extraction.

4.6 Error Analysis

We perform error analysis after experiments. A typical error is 姜文 ‘Jiang Wen’, a famous movie actor in China which is wrongly recognized as a topical key concept. The vertices that closely related to 姜文 ‘Jiang Wen’ in the graph are 宁静‘Jing Ning’, 夏雨‘Yu Xia’ and 阳光灿烂的日子‘In the Heat of the Sun’. We believe that the best topical key concept about this topic is 文艺 ‘literature’. However, ‘literature’ cannot be created if it never appears in the tags of Douban.com. This error is due to randomness of folksonomy tagging itself. We consider integrating other relative folksonomy data sources such as Baidu video[3] to overcome this defect in the future work.

5 Conclusion

In this paper we study the novel problem of topical key concept extraction from folksonomy. A new link weight function is proposed to improve graph-based ranking method for topical key concept extraction. Quantitative and qualitative evaluations indicate the robustness and effectiveness of our method. In the future, we will make full use of the topical key concepts and relevant entities, and also the relationships byproduct for Chinese ontology construction. Experiments on the folksonomy data from Douban.com Movie show that our method is feasible. We will further explore our method in other domains such as music and more large-scale data with the help of other folksonomy-based systems.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) via grant 61273321, 61133012 and the National 863 Leading Technology Research Project via grant 2012AA011102. Finally, we would like to thank the anonymous reviewers for their insightful comments.

References

Angeletou S, Sabou M, Motta E. 2008. Semantically Enriching Folksonomies with FLOR. In *the 1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web*, Tenerife, Spain, pages 1-16.

Blei D M, Ng A Y, Jordan M I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, volume 3, pages 993-1022.

Buitelaar P, Cimiano P, Magnini B. 2005. *Ontology Learning from Text: Methods, Applications and Evaluation*, volume 123, pages 3-12.

Cui G, Lu Q, Li W, et al. 2009. Automatic acquisition of attributes for ontology construction. In *Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy*, Springer Berlin Heidelberg, pages 248-259.

Haveliwala T H. Topic-sensitive pagerank. 2002. In *Proceedings of the 11th Association for Computing Machinery international conference on World Wide Web (ACM)*, pages 517-526.

Hearst M A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Association for Computational Linguistics conference on Computational linguistics-Volume 2 (ACL)*, pages 539-545.

Hotho A, Jäschke R, Schmitz C, et al. 2006. Information retrieval in folksonomies: Search and rank-

[3] http://video.baidu.com

ing. *The Semantic Web: Research and Applications*, Springer Berlin Heidelberg, pages 411-426.

Hulth A. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Association for Computational Linguistics conference on Empirical methods in natural language processing (ACL)*, pages 216-223.

Jin Y, Li R, Wen K, et al. 2011. Topic-based ranking in Folksonomy via probabilistic model. *Journal of Artificial Intelligence Review*, 36(2), pages 139-151.

Liu X, Song Y, Liu S, et al. 2012. Automatic taxonomy construction from keywords. In *Proceedings of the 18th Association for Computing Machinery international conference on Knowledge discovery and data mining (ACM SIGKDD)*, pages 1433-1441.

Liu Z, Huang W, Zheng Y, et al. 2010. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the Association for Computational Linguistics Conference on Empirical Methods in Natural Language Processing (ACL)*, pages 366-376.

Marlow C, Naaman M, Boyd D, et al. 2006. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of the 17th Association for Computing Machinery conference on Hypertext and hypermedia (ACM)*, pages 31-40.

Mihalcea R, Tarau P. 2004. TextRank: Bringing order into texts. In *Proceedings of Association for Computational Linguistics Conference on Empirical Methods in Natural Language Processing, 4(4)*, pages 404-411.

Nakayama K, Pei M, Erdmann M, et al. 2008. Wikipedia Mining-Wikipedia as a Corpus for Knowledge Extraction. In *Proceedings of Annual Wikipedia Conference*, pages 1-15.

Nie L, Davison B D, Qi X. 2006. Topical link analysis for web search. In *Proceedings of the 29th annual international Association for Computing Machinery conference on Research and development in information retrieval (ACM SIGIR)*, pages 91-98.

Page L, Brin S, Motwani R, Winograd T. 1999. The pagerank citation ranking: Bringing order to the web. *Technical report, Stanford Digital Library Technologies Project*, pages 1-17.

Trant J. 2009. Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information*, 10(1), pages 1-42.

Voorhees E M. The TREC-8 question answering track report. 1999. In *Proceedings of TREC*, pages 77-82.

Voorhees E, Harman D, Standards N I, et al. 2005. TREC: Experiment and evaluation in information retrieval . *Cambridge: MIT press Boston*, pages 1-567.

Zhao X, Jiang J, He J, et al. 2011. Topical keyphrase extraction from Twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL)*, pages 1-10.

Zhou M, Bao S, Wu X, et al. 2007. An unsupervised model for exploring hierarchical semantics from social annotations. *Journal of the Semantic Web*, pages 680-693.