Semantic Relations Mining in Social Tags Based on a Modern Chinese Semantic Dictionary

Jiangying YU [0000-0003-1822-1461]

Yunnan Open University, Kunming, Yunnan, 650223, China
Beijing Language and Culture University, Beijing, 10083, China
yujiangying2008@qq.com

Abstract. At present, many scholars have studied the semantic relations mining in social tags based on *WordNet*—an English semantic dictionary and have made some progress. There have been few studies to combine modern Chinese semantic dictionary and social tags. The paper selects tag data from *Dòubàn Reading* first, then uses the classification and coding system of *A Thesaurus of Modern Chinese*(TMC), calculates the semantic similarity of tag data and mines the semantic relations in social tags by *WordSimilarity*—a lexical semantic similarity computing system. The results obtained with this method, not so different from the way we think of lexical semantic relations, have a higher accuracy.

Keywords: Semantic Relations, Social Tags, Semantic Dictionary.

1 Introduction

In recent years, with the development and popularization of Web2.0 such as del.icio.us, flicker, Dòubàn, etc, the generation, organization, publishing and sharing of Internet information have been changed, network users have become more and more important and a user-centered social network has gradually formed. Network information users spontaneously choose proper words to describe certain types of resources according to their own understanding of information resources. This free-optional, convenient and flexible way of classification has been welcomed by the network information users, since the markup language is not subject to any restrictions. For folksonomies' labels, on the other hand, there are some shortcomings; for instance, the diversity, fuzziness, and unorganized state of the labels and the lack of semantic relations between words, which not only seriously affect the efficiency of information retrieval, but are difficult to adapt to the requirements of the semantic web. Therefore, we hope to optimize the folksonomy system, improve the efficiency of the network information dissemination and retrieval and construct a semantic network between tags.

Literature research reveals that empirical studies on specialized systems for Chinese characteristics have not been carried out yet in China. A small amount of empirical data basically comes from foreign popular folksonomy websites such as

Y. Wu et al. (Eds.): CLSW 2017, LNAI 10709, pp. 315–320, 2018.
https://doi.org/10.1007/978-3-319-73573-3_28

del.icio.us, and instances are also selected from foreign online dictionary such as WordNet. As Chinese language has its own characteristics, transplanting the results of foreign research simply is not effective. Therefore, this paper tries to mine the semantic relations in social tags by selecting tag data from *Dòubàn Reading* and *TMC*.

2 A Thesaurus of Modern Chinese (TMC)

SU Xin-chun and his team have completed *A Thesaurus of Modern Chinese(TMC)* in 2013. *TMC* inherits the tradition of conceptual classification since *Synonym Dictionary* to reflect conception relations of the whole society and human recognition. It embodies more than 80,000 modern Chinese words with high frequency and constructs a five-level semantic classification system with 9 classes in the first level, 62 in the second level, 508 in the third level, 2,057 in the fourth level and 12,659 fifth-level classifications. This kind of semantic classification emphasizes the governing function from upper semantic levels to subordinate levels, the coverage function of the subordinate semantic levels to the upper levels and the complementary function between the neighboring semantic levels.

Different methods have been respectively used in the five-level semantic classification. Numbers in capital Chinese characters are first-level, lower-case characters are second-level, capital letters are third-level, lower letters are fourth-level, Arabic numerals are fifth-level. Therefore, you can clearly show the semantic class hierarchy and sequence with a set of numbers, and every word has a unique "ID." For example, the word "people" has an ID of "壹一Aa01," while the word "dictionary" has an ID of "叁八Eb29."

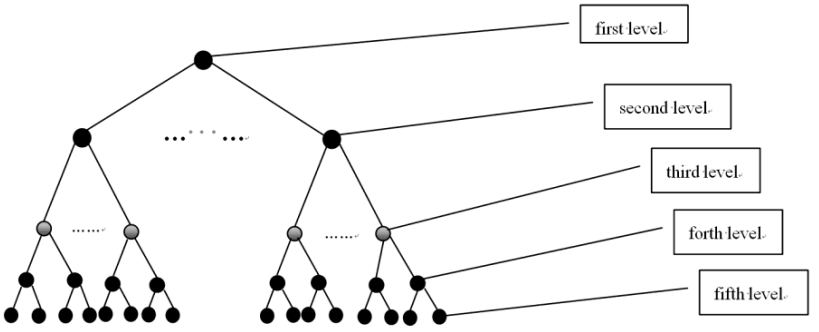

Fig. Five-level semantic classification system of *TMC*.

3 Word Similarity Computation Based on *TMC*

Polysemy is prevalent in the Chinese language. One word may have more than one meaning. This is what we call polysemy, and such a word is called a polysemous word. Multiple meanings of a polysemy are usually placed under the same item of one word in a traditional dictionary, while different meanings of a polysemy are placed

under different items in *TMC*. The word "milk," for example, has three IDs: 壹五 Ae08, 壹五 Ca07, and 伍七 Cd02. Therefore, when calculating the similarity of Chinese words, we should consider all the meanings of one word. When calculating the similarity of two words, we first find out different IDs for different meanings of polysemy in *TMC*, then calculate its similarity, and finally take the maximum of the similarity as the similarity of two words. The specific calculation method is to first judge two meanings of polysemy as a leaf node branch from which level in *TMC*, and see the IDs of the two meanings of polysemy differ from which level. Judging from first-level, if IDs are kept the same, multiply this number by 1, otherwise, multiply by the corresponding coefficient from the branch level. In addition, to guarantee the similarity of two meanings' controls in [0, 1], multiply it by a parameter adjustment $\cos\left(n * \frac{\pi}{180}\right)$, then multiply it by a controls parameter $\frac{n-k+1}{n}$ since the similarity will be directly affected by it to branch from which level. Among them, n is the total number of nodes of branching, k is the distance between two branches, so you can get more precise similarity of different meanings of polysemy.

In addition, we also introduce a word semantic similarity computing system called *WordSimilarity*. The system can calculate the semantic similarity between more than one word at the same time. We will choose this system to calculate the semantic similarity between words tags, which generally fall into two cases:

First, simple words, which have been included in *TMC*. We can directly calculate the semantic similarity between simple words with the *WordSimilarity* system, which can also be used as the credibility of the corresponding matching concept. For instance, the semantic similarity between "literature" and "novel" we calculated is 0.899.

Second, compound words, which are formed by some simple words and have not been included in *TMC*. For compound words, firstly, we complete the compound word segmentation by positive maximum matching method, then calculate the semantic similarity between simple words after the segmentation by *WordSimilarity* system, and finally we take the average value as the semantic similarity between compound words. For example, we calculate the semantic similarity between "literature " and "Chinese literature" like this: firstly, as a compound word, "Chinese literature" has been divided into "Chinese" and "literature," then we get the semantic similarity between "literature" and "Chinese" which is 0.899, then between "Chinese" and "China" which is 1.0 according to the *WordSimilarity* system, and finally we take the average value as the semantic similarity between "literature" and "Chinese literature" which is 0.793.

4 Empirical Studies

4.1 Tag Data

Douban.com (Chinese: 豆瓣; pinyin: Dòubàn), launched on March 6, 2005, is a Chinese SNS website allowing registered users to record information and create content

related to film, books, music, and recent events and activities in Chinese cities. It can be seen as one of the most influential web 2.0 websites in China. Unlike Facebook and Renren, *Dòubàn* is open to both registered and unregistered users. For registered users, the site recommends potentially interesting books, movies, and music to them in addition to serving as a social network website and record keeper; for unregistered users, the site is a place to find ratings and reviews of said media.

In order to conduct research, we first extract the top 5 books from the 2016 list of Chinese literature in *Dòubàn Reading*, then extract the five *Dòubàn* members of the most respectively commonly used tags from these books resource tags, and finally we get a sample data set containing 25 labels (see Table 1).

Table 1. Label sample data set from *Dòubàn reading*.

book name	tag 1	tag 2	tag 3	tag 4	tag 5
午夜起来听寂静	poetry	poem	Chinese Literature	Literature	China
重读	essays	Literature	prose	Chinese Literature	writing
走进一座圣殿	essays	life	prose	Literature	China
平原上的摩西	novel	Chinese Literature	China	short story	Literature
台北人	novel	Chinese Literature	short story	Literature	China

4.2 Semantic Relations Mining in Social Tags Based on TMC

As before, we use *WordSimilarity* to calculate the semantic similarity between different tag words. Of these tags, the results for the simple words "poetry, poem, literature, Chinese literature, China, essays, prose, novel, short story, writing, life" can be found in Table 2 below.

Table 2. Semantic similarity between tag words based on *WordSimilarity*.

poetry	1.000										
poem	1.000	1.000									
Chinese	0.622	0.622	1.000								
Literature	0.657	0.657	0.793	1.000							
China	0.586	0.586	0.793	0.586	1.000						
essays	0.678	0.678	0.632	0.678	0.586	1.000					
prose	0.678	0.678	0.632	0.678	0.586	0.765	1.000				
writing	0.100	0.100	0.100	0.100	0.100	0.100	0.100	1.000			
life	0.586	0.586	0.451	0.586	0.315	0.586	0.586	0.541	1.000		
novel	0.657	0.657	0.743	0.899	0.586	0.678	0.678	0.100	0.586	1.000	
short story	0.657	0.657	0.743	0.899	0.586	0.678	0.678	0.100	0.586	0.959	1.000
	poetry	poem	Chinese	Literature	China	essays	prose	writing	life	novel	short story

We can see from the above calculation results:

(1) The semantic similarity between the two labels words "poetry" and "poem" is 1. Then we check the result in TMC and find that the ID of poetry and poem is the

same (叁八Dd01). Thus, it could be concluded that "poetry" is synonymous with "poem.".

(2) There are several pairs of labels words whose semantic similarity is greater than 0.65 and less than 1 (< literature, poetry/poem >, < literature, Chinese literature >, < literature, essays>, < literature, prose >, < literature, novel >, < the literature, the short story >, < China, Chinese literature >, < essays, poetry / poem >, < essays, fiction >, < essays, short stories >, < prose, poetry / poem >, < prose, novels, essays, short stories >, < novels, poetry / poem >, < novels, Chinese literature >, < short stories, poetry / poem >, < short stories, Chinese literature >, < short story, novel >). These tag words share a high semantic similarity and constitute near-synonymy units.

(3) The semantic similarity between the tag word "writing" and other tag words is 0.1. That is to say, there is no correlation between the tag word "writing" and the other tag words. Then we check the result in TMC and find that the ID of "writing" is 陆五Ea01, which is viewed as clearly different from IDs of other tag words. Therefore, it can be concluded that there are no semantic relations between them.

5 Conclusions

Folksonomy is an important way of information organization in the network age. Semantic relations in mining through social tags can both greatly optimize the classification system of the masses, and provide theoretical support for the next generation of Internet comprehensive implementation. This paper is only a trial study with small sample sizes on semantic relations mining and as the experimental results show, the semantic relations mining based on *TMC* has a good accuracy. It's important to note that the inadequacies of the study are mainly the limitations regarding the research tools and sampling, limiting us from studying the subject further.

References

1. Chklovski, T., Pantel, P.: VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In: Proceedings of Conference on Empirical Methods in Natural Language Processing(EMNLP-04), pp. 33-40. Barcelona, Spain (2004).
2. Miller, G. A.: Wordnet: A Lexical Database for English. Communications of the ACM 38(11), 39–41(1995).
3. Dòubàn Homepage, https://en.wikipedia.org/wiki/Dòubàn, last accessed 2017/03/25.
4. Feng, L. I., Fang, L. I.: An New Approach Measuring Semantic Similarity in Hownet 2000. Journal of Chinese Information Processing21(3), 99-105 (2007).
5. Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., & Stumme, G. Evaluating Similarity Measures for Emergent Semantics of Social Tagging. In: Proceedings of the International Conference on World Wide Web, pp.641–650. New York, NY, USA (2009).
6. Melnik, S., Garcia-Molina, H., & Rahm, E. Similarity Flooding: A Versatile Graph Matching Algorithm. In: Proceedings of the 18th International Conference on Data Engineering(ICDA), pp.117–128. Santa Barbara, CA, USA (2002).

7. Madhavan, J., Bernstein, P. A., & Rahm, e on Very Large Data Bases, pp.49–58. Morgan Kaufmann Publishers Inc, Italy (2001).
8. Su, X. CH. A Thesaurus of Modern Chinese (TMC). 1st edn. The Commercial Press, Beiji ng, CHN (2013).
9. Xiong, H. X.: Research Overview on the Combination of Tag and Ontology in Social Tagg ing System. Journal of Intelligence 32(8), 136–141 (2013).
10. Xiong, H. X., & Wang, X. D.: Research on Mapping between Tag and Ontology in Folk- sonomy System. Information Science32(3), 121–126 (2014).

Supported by the National Language Committee of China (Grant No. YB125-170).